Couleur rapidement donc pour les signatures, ça c'est vous qui gérez, moi j'ai rien à faire en termes d'organisation. Une note écrite quoi à la fin, donc ça. Le sujet est pas encore fait, mais je dirais en fonction de ce qu'on arrive à faire ensemble, et cetera, et il y aura une note de TP, donc ça ce sera 1TP qu'on vous donnera à la dernière séance. Et en avance. Donc si vous voulez commencer avant, vous pourrez juste sur la dernière séance, on sera là ensemble pour t'aider. Et vous serez ce noté sur un donc le code et le rapport qui a été géré avec sur ce petit projet là et donc au cours de ce cours donc on va avoir là on a une séance uniquement magistrat en quelque sorte. Donc une position des bases de manière très vite comme ça il y a beaucoup beaucoup de notions en renforcement de Farming et c'est important de bien noter toutes les définitions. Quand je parlerai de value, fonction de gain de ce que vous voulez, il faut absolument savoir de quoi on parle. Oui. Ouais c'est ça, il y aura un TP noté sur la dernière séance et en DS plus tard sa je sais pas on verra. On sera mieux. Ouais, je veux dire c'est il y a des notions très spécifiques et c'est important de donner. Donc. Et pour le tp ? Il y a plusieurs petites épées, je vais vous mettre pendant les séances. Si vous voulez être efficace sur le P final, Bah c'est important de bien faire les TP au milieu des séances ils sont pas notés vous verrez, ils sont très simples mais ça me permet de bah de d'arriver à maîtriser un peu les librairies qu'on utilisait. Ok, on va pas recommencer vu que les derniers sont arrivés. Alors très rapidement le programme de secours pour l'instant, sachant que ça va être modulable en fonction aussi des des sondages à faire. Là cette séance là on va faire une introduction générale à qu'est ce que c'est que le renforcement learning ? Et je vais vous poser les les bases mathématiques mathématiques c'est un gros mot mais vous Notez toutes les définitions de les thèmes qui retrouver tout le temps de cette séance. Là elle est très importante et je vais je vais vous noter à chaque fois les définitions qu'il faut bien qu'il faut bien retenir. Et donc ça vraiment, c'est c'est des trucs limites à connaître par cœur ou dire à vraiment à bien maîtriser en termes de mathématiques. Et donc le cas de mathématiques du RL, c'est les marqueurs des process. A un petit point, ce cours ce sera en français. Tout le monde en parle français. Mais les salles sont en anglais. C'est pour des raisons pratiques et aussi parce que en en en R on utilise tout le temps des termes anglais. Et en fait quand vous allez lire le littérature ou voir du code, ce sera tout le temps en anglais. Donc je pense que c'est mieux de pas traduire les termes, les termes techniques. Bonjour. Donc les termes techniques resteront en anglais, je les traduirai de temps en temps mais en tout cas sur les failles ils seront en anglais. Voilà bon ensuite pour le reste ça, surtout quand vous aurez les slides pour vous repérer sur moi ce qu'on est encore. Donc. Pour commencer ? Rapide, très rapide donc point les les manières qu'on a d'apprendre des modèles avec de la donnée. On a l'apprentissage le plus classique et l'apprentissage supervisé. OK. Apprentissage supervisé, on sait quelle quelle classe on a pour chaque jeu de données à dire. On sait que mon point là il a été classifié dans ta class et cetera, et on a le modèle. Et à chaque fois que le fait une prédiction on peut dire Ah ça c'est pas bien, il fallait prédire la classe B. Ça c'est super lol. Ensuite on a de de l'apprentissage non supervisé, non supervisé. C'est quant vous avez avoir plein d'images et vous allez demander au modèle à ce que lui même il va. Il va faire les différences entre ce qu'il voit sans pour autant qu'il y ait de des classes. Donc typiquement pour ceux qui suivent un peu l'actualité, les modèles type. Type les clips, vlm, et cetera. On a donc des modèles qui vont voir beaucoup, beaucoup d'images, mais toutes les images ne sont pas labellisées. En fait, ce qu'on va dire au modèle c'est en fait je vais te donner plein d'images, un peu comme nous, comme on apprend, comme des êtres humains, on voit plein d'images et on fait la différence entre les choses. On se dit Ah bah tiens ça ce ce truc là c'est une roche, connais pas. Et donc là ensuite on va, on va peut-être demander, on va mettre une nouvelle classe. Donc c'est pour ça que ici le clustering c'est un truc de d'apprendre à se spécialiser parce que là on voit que voit que entrer et traduction, on a un modèle qui va interagir avec un environnement, donc c'est un ça devient un agent. On a un environnement, on a un agent qui évolue dans cet environnement et on va essayer de faire en sorte que notre agent, il fasse des choses intéressantes dans ton environnement. Et donc ce qui est compliqué avec le renforcement learning, c'est que on n'a pas forcément. Un retour pour le modèle à chaque à chaque fois qu'il fait une action. Typiquement, si vous jouez aux échecs, on peut je peux pas vous dire facilement au moment où je où je place mon cavalier en D 4, vous pouvez pas me dire Ah ça c'est super ou parce que vous êtes fort. Vous pourrez peut-être me le dire mais en fait ce qui compte c'est qu'à la fin je gagne. Donc en fait le modèle il sait pas, il sait pas ce qu'il fait jusqu'au dernier moment où il gagne ou il gagne pas. On dit Ah c'est bien ou c'est pas bien ? Ça veut dire qu'au moment où il gagne, va dire Ah mais j'ai gagné, donc ça veut dire que peut-être que le 3e coup que j'ai fait au tout début de la partie, c'était un super coup. Mais donc en fait parfois pour donner un peu de pourquoi, c'est beaucoup plus difficile de d'apprendre dans le cas de du retranchement. Euh je sais pas, vous pouvez rentrer par la porte de derrière après si vous êtes en c'est pas. Donc ouais donc là on revient sur le l'apprentissage supervisé. Entrez, on a notre notre sortie, on a les targets et on prédit la la valeur. Donc un petit même rigolo, mais c'est vraiment à chaque fois que le modèle dit quelque chose on va dire c'est bien, c'est pas bien. Donc le renforcement de learning. On a un objectif pour le modèle mais on. On n'a pas forcément de récompense tout de suite. Et il y a un autre problème que que vous expose et avec cet outil slide, c'est que en rénovation learning, souvent avoir un modèle de récompense pour dire au modèle c'est bien ou c'est pas bien, mais en fait, qu'est-ce qu'on va lui dire ? On va lui dire on va le donner. Une récompense qui va être chiffrée, c'est-à-dire qu'elle va pouvoir être plus ou moins bonne. La récompense. Et donc après y a tout un problème de. Bah en fait quand on lui donne moins 10, bah ça veut dire quoi ? Ça se trouve la récompense minimale qui peut avoir c'est moins 1000 donc en fait moins 10 il s'en fout quoi. Donc il y a un problème aussi donc dont on va parler qui est un peu de de de l'exploration, c'est à dire que le modèle va devoir explorer par lui-même un peu le le le monde extérieur pour se rendre compte que. Bah en fait. Vous dans dans vas-y tous les jours si vous avez leur tête pour votre conduite, c'est c'est c'est pas bien. La prochaine fois vous le referez pas mais c'est pas très grave. Alors si vous tombez de 2 étages, ça commence à être un peu plus compliqué. Il y a une nécessité d'exploration. Hop. Des petits exemples de de fonction cours qu'on peut qu'on peut utiliser, donc de rewards qu'on peut utiliser si on veut apprendre à un hélicoptère, donc un drone à à faire des à faire des des manœuvres, on va lui dire, on va lui donner une trajectoire et si le le drone il suit la bonne. Trajectoire. Alors on lui donne une récompense et plus il s'éloigne de la trajectoire, plus on va lui lui enlever de la réponse. Là c'est le terme de plein dans le ciel. Mais je pense que je vais utiliser cela directement. Je. Donc reward, penalty facile, plutôt clair. Pour des modèles de R qui font de l'investissement en finance, on va bien entendu indexer leurs récompenses sur les sous qui font gagner à la boîte, ça c'est facile. Pour donc une, une station, une station de d'électricité, Ben pareil on va. On va faire en sorte de produire un maximum tout en tout en minimisant les coûts de production. Donc je sais pas si vous avez une usine qui va produire de l'hydrogène avec des des circuits d'électricité du du réseau, et bah on il va apprendre lui-même. Il y a des modèles chez EDF et chez RTE qui aujourd'hui ? Du rayon force motoring pour optimiser la production. Donc typiquement, soit on va stocker dans des. Le barrage dans des batteries, et cetera. Et il y a des modèles de renforcement learning qui apprennent en fonction de ce qu'ils voient sur le réseau. Ah Ben en fait là il faut stocker, là il faut produire. Donc ils apprennent ça avec des méthodes de R. Si vous suivez un peu les actualités sur les les nouveaux robots humanoïdes open Aya figure y A quoi qu'on qu'on sort des trucs en ce moment et à suivre le Boston Dynamics qui est très chouette aussi. Eh Ben on va aussi avoir des fonctions de coûts du type Ben en fait, si le robot il tient debout, déjà c'est une bonne chose. Je vous montrerai un exemple tout à l'heure et on va le dire, Bah c'est bien de se tenir debout, puis ensuite c'est bien d'avancer, c'est bien de de se rapprocher de de cela, et cetera. Voilà bon pour des pour des pour des jeux vidéo c'est pareil, vous allez mettre une récompense qui gagne. D'accord ? On va commencer. Donc là, c'est le moment où vous avez commencé à prendre des notes ? On. On va en poser petit à petit. D'abord, je vais poser les les définitions intuitives de toutes les notions qu'on va utiliser et ensuite on ira plus en détail sur les définitions mathématiques. Donc on se place dans le cadre du renforce. Learning. On a un environnement, un agent et on a, on va avoir une interaction. On veut faire en sorte que l'agent se déplace, fasse des choses dans l'environnement. La première chose qu'on reçoit quand on quand on on est pas dans l'événement, c'est une observation, c'est-à-dire l'environnement. Donc en pratique dans le code vous allez avoir un une classe environnement. Une classe agent, il y aura des communications entre les 2. La première chose qui va recevoir votre agent c'est une observation ça ? Je pense que vous avez déjà dû le faire rapidement. Donc quand vous voyez 1O c'est une observation. L'observation, elle peut être multiple dans le cadre d'un jeu Atari. Typiquement, vous allez voir le titre qu'on va donner au modèle, l'écran de votre PC et on va demander au modèle d'appuyer sur les touches du clavier. Donc l'observation ce sera les pixels de votre écran. Donc c'est ça c'est simple. Ensuite, on va avoir la reward, la récompense R. Elle c'est pareil. Donc là c'est ce qu'on a vu. Y a plein de de modèles de reward qui peuvent être différents, mais c'est ce qui va permettre à l'agent de comprendre si ce qui est en train de faire c'est bien ou pas. Et ensuite on a l'action, donc ça c'est ce que l'agent va renvoyer à l'environnement. Donc là c'est un peu plus particulier parce que l'idée qu'on se fait d'une force avec, là ici c'est des exemples, c'est un bébé. C'est le bébé qui fait tout seul ses actions. Dans le cadre du RL où on est on avec des programmes informatiques en fait, on va avoir le modèle qui va choisir quelle action faire, mais c'est l'environnement qui va calculer l'impact de son action. Est-ce que c'est clair ? Donc par exemple, si. Si vous avez un. Un un modèle de robot, le modèle de robot ? Il va voir l'environnement, il va voir tous ces censeurs, donc il va me dire si le remettre il va avoir des des capteurs un peu partout et il va se dire OK là je vais choisir de lever ma jambe gauche donc il dit OK je vais activer le moteur va se mettre jambe en plus. Oups. Et donc il va envoyer ça au modèle de au modèle de du monde parce qu'on est on est en simulation hein ? Le robot est pas réel pour le moment et le le modèle du monde va dire OK il active son moteur, qu'est-ce qui se passe ? Bah la jambe va se lever, peut-être qu'il va tomber parce qu'il était en déséquilibre. Voilà et là c'est le le l'environnement qui va calculer ce qui se passe ensuite et il va ensuite répondre à l'argent. Il va faire OK ? T'as fait telle action et Ben il se passe ça maintenant, donc tu te retrouves dans telle nouvelle situation. Et la situation, c'est le l'État. Ouais. Euh. Ah bah c'est ouais en fait j'avais j'avais défini déjà tout ça. Alors qu'est-ce que je dis ? Donc ça, c'est pour l'observation. Ouais, ensuite. Donc pour ça c'est bon pour pour l'observation, c'est clair pour tout le monde. Ouais, ouais. Oui. Ce qui se passe, c'est que. On on va y revenir dessus. Mais le cas de mathématiques pour décrire ce cette interaction entre je reçois une observation j'émets une action, il se passe un nouvel État et je reçois une récompense. Il se décrit dans le cas de mathématiques des des processus de décision Markoviens au Markov décision Process et donc dans ce cas-là c'est un un punchy observer. Et c'est le fait que tu peux pas tout observer. Voilà donc ça c'est franchement c'est c'est pas très, c'est pas très important, mais garder le MDP on va, on va en reparler tout à l'heure mais c'est pas grave, on va aller plus en détail là. Le but c'est vraiment que vous ayez une intuition sur c'est quoi qu'on manipule ? C'est quoi les variables ? Qu'on pose quoi ? Donc l'agent, c'est le notre système qui qui va faire nos actions, on a l'environnement. On a l'action. Vous voyez pas là, d'accord, c'est là. Ouais, pour la température non ? Voit que non, c'est pas. On était à l'action donc l'action a c'est bon pour tout le monde. L'action c'est quoi ? Qu'est ce que j'active et quelle action j'envoie à l'environnement ? Ensuite voilà l'État, donc l'État qui sera noté s c'est pour l'environnement, c'est pour qu'il sache dans quel État est notre système. Donc dans le cas d'un jeu Atari ça va être je sais pas si vous êtes sur le un un modèle qui veut sortir de labyrinthe là ou si vous frodon live typiquement. L'état de mon système, c'est où est-ce que se trouve mon bonhomme potentiellement ? Ou ce qui se trouve dans dans quelle direction ils regardent, à quel objet il a dans ses poches ? J'en sais rien. L'état de notre système. Et la, la reward, donc le feedback qu'on donne à notre agent. Alors là, on a une petite vidéo. Donc je sais pas si vous connaissez cette chaîne Youtube des ça c'est assez marrant mais ça vous donne un peu, ça vous donne de très bonnes intuitions sur comment est-ce que en pratique on fait pour entraîner des modèles ? Et donc là dans cette vidéo bah laissez vous le présenter. On va, on va, on va essayer de faire apprendre un robot à marcher. Donc en pratique c'est que de la simulation, mais c'est ce genre d'algorithmes qui sont aussi utilisés pour vraiment. Ensuite on va pouvoir uploader le modèle d'i a dans un vrai robot. Donc là, typiquement, on a un cas. On a un cas où on a. On a mis une fonction coût qui est très simple pour le modèle. Vous avez compris si intéressé ou pas ? Ouais donc là le but c'est d'apprendre au modèle à marcher. Et du coup on s'est dit bah c'est facile, quelle récompense on va mettre à notre modèle pour que il apprenne à marcher, on lui dise facile va là-bas. Sauf que le modèle il sait même pas comment tenir debout en fait. Donc il va faire tout ce qu'il peut-il va bouger tous ses membres parce qu'il a accès, donc il peut bouger ses bras, ses jambes, ce que vous voulez, il va faire tout ce qu'il peut pour aller là-bas, mais il a aucune notion de de savoir marcher en fait. Donc là ce qui se passe c'est qu'il a appris à ramper quoi. Hop hop, chut. Ouais donc là notre mot il a appris à ramper. Donc là, ce qu'on va faire, c'est que on va. On va, on va augmenter à discuter pour faire en sorte que le modèle, il ne puisse plus utiliser cette technique de de camping et qu'il soit obligé d'apprendre à marcher. Donc là ça marche toujours pas. Et donc maintenant, ce qu'on pêche parce que vous voyez, mais on va aussi le punir pour avoir ces membres qui touchent le sol. C'est en pratique un peu comme un être humain qui va être puni à si il tombe, il va avoir mal quoi. Et donc il va essayer de se mettre debout. Donc là voilà donc là, typiquement, on commence, on commence à arriver à se tenir debout. Donc on continue, on continue, on continue, on arrive toujours pas. On commence à avoir quelque chose qui ressemble à de la marche. Mais bon, on a encore des choses à apprendre. Vous voyez aussi que il y a beaucoup, beaucoup, beaucoup d'essais, donc beaucoup de simulations successives pour que le modèle il commence à apprendre des choses. Ça c'est un truc particulier, une force planning. Ça coûte très cher parce que vous avez beaucoup d'itérations en général pour. Pour arriver à un bon résultat. Oui. Comment on décide de réinitialiser l'État ? Parce qu'il y a un moment-là il est bloqué donc on le réinitialise même. Ouais bah là c'est c'est à toi de choisir. Alors typiquement quand on agent il meurt, quand t'es dans un jeu vidéo, quand tu perds la partie, quand t'es dans une impasse. Financier ? On parlait tout à l'heure, ouais. Non en général en fait tu vas recommencer dans tous les cas. Donc typiquement ce qu'on va faire c'est que là on le voit pas, mais tu vas le faire par batch, c'est à dire que t'as le même modèle, tu vas l'entraîner 10 fois en parallèle. En gros tu vas récupérer, tu vas faire 10 essais et en gros sur les 10 essais tu vas dire Ben en fait il aurait bien fallu en laisser un, tu remarques en fait il fallait bien faire telle action à tel moment. On essaie 2 tel autre, et cetera, et ensuite tu vas rassembler tout ça, faire ta tambouille donc tu vas en gros tu fais un pour ceux. Du coup vous avez déjà fait un peu de de. De \*\*\*\*\* et compagnie non ? Donc en gros vous calculez vos gradients surtout sur tout votre batch. Vous faites une moyenne et vous faites votre update de votre modèle donc en pratique invasion de neurones petit à petit voilà et donc dans le cadre du modèle financier, typiquement, tu vas choisir de. De faire une prédiction sur uniquement 100 100 ? Pas de temps et c'est pas grave. Tu t'en fiches de voir à quel point le modèle il est bon. Tu vois peut-être que dans certains cas si le modèle est trop trop mauvais, tu vas en effet tu vas juste écarter cet épisode de ton de ton entraînement. Mais en pratique en pratique ça change pas grand chose. Donc là on en est où alors le modèle ? Donc là je je passe, je crois que c'est il arrive à marcher mais il arrive pas à tourner c'est ça ? Donc là maintenant donc le CE que enfin le franchement le le mec qui programme ça est très bon parce que il à chaque fois qu'il a un nouveau problème, il arrive à trouver une nouvelle situation dans laquelle le modèle va pouvoir apprendre ce qui ce qu'il attend de lui. Donc typiquement là on arrivait pas, on n'arrivait plus à tourner. Eh Ben on on fait un nouvel épisode où on a des murs donc le modèle obligé de tourner et on a une on a une une récompense successive au fur à mesure. Donc le modèle est obligé de savoir tourner pour pouvoir pour pouvoir apprendre. On va passer un petit peu. Donc là donc donc le modèle apprend à le modèle apprend à à marcher et à tourner. Et maintenant ce qui se fait aussi beaucoup en repensant learning, c'est on va le soumettre à des à des des perturbations. Ça permet que le modèle soit robuste, pas typiquement votre modèle s'il apprend à marcher sur un sol parfaitement plat. En fait le jour où y aura une toute petite aspérité sur le sol et bah il va comme il a jamais vu cette situation là et bah il va tomber tout de suite. Donc on on entraîne les modèles avec ce genre de ce genre de méthode, on lui balance des cubes à la gueule et et on le modèle apprend très bien à garder la, à garder son équilibre. Hop. Allez, je vous laisse, je vous laisse le. Donc une fois qu'il a appris toutes ces. Toutes ces petits skills à droite à gauche, à la fin on l'entraîne sur. Sur un épisode final, sur un niveau final où on va rassembler un peu toutes ces compétences là pour vérifier que le modèle il a pas des appris typiquement à rester debout, des appris à tourner et cetera, en évitant les blocs ou en faisant. Donc là on voit que il arrive bien à faire ses pas. Ensuite, on va essayer de le faire tourner. Oui bon, il marche un peu comme un constipé, mais. Et en pratique, c'est vraiment ce genre de choses qui sont utilisées pour apprendre à des des vrais robots à marcher. Et, et c'est notamment les méthodes de de perturbation qui permettent de de comment dire de de passer d'un modèle uniquement en virtuel à un modèle qui peut être appliqué dans la vraie vie. Donc là on arrive sur le passage des billets cubes. Et on y est presque. Et voilà. Vous avez envie de faire pareil ? Ok. Donc ça, c'est un cas pratique. Comme vous le voyez, c'est très difficile d'entraîner ce genre de ce genre de robot. Vous allez voir donc sur le dernier TP on aura enfin on on envisage c'est on en reparlera si vous voulez faire autre chose. Mais faire apprendre une voiture à évoluer sur un sur une autoroute donc pas pas se taper les autres voitures. De trouver son chemin, savoir dépasser et cetera. Et et donc c'est ce que vous allez devoir faire ce genre de choses, c'est à dire il suffit pas de dire à la voiture d'aller vite parce que il va falloir lui mettre un une récompense négative contestable une autre voiture, mais pas trop non plus parce que sinon en fait si vous mettez une récompense trop négative, bah en fait elle va juste s'arrêter, donc elle va être elle ne tape pas. Ah ne tapera désolé, plus de voiture, mais elle va plus avancer non plus. Donc vous avez tout ce genre de de petite, de petites choses à mettre en place pour que vous. Votre modèle apprenne à à faire des choses tout en restant en sécurité, et cetera. Bon, maintenant qu'on sait un peu de quoi on parle. On va ? On va corser un peu les choses. En posant +1 peu plus les les, les définitions mathématiques, donc nouveau terme, le terme de politique pour aussi. La politique c'est en fait c'est notre modèle. Donc le pile hein, c'est le modèle qui va dire à partir de mon État. Donc pour le modèle si je suis dans l'état s ce que je fais. Donc typiquement, on va s'appuyer beaucoup sur le cas du sur un cas très simple qui est celui du labyrinthe. Ok. Le cas d'une adhérence. C'est facile, on est dans un labyrinthe. Notre État c'est là où on se trouve, dans le labyrinthe des coordonnées et l'action. Donc l'action qu'on va donner, c'est dans quel cas je bouge parce que je vais à droite, je vais à gauche ce que je vais en haut, en bas, sachant qu'il y a des États qui sont impossibles. Donc quand je suis ici, je peux pas aller à gauche. Typique. Et donc le cas du labyrinthe. En fait, c'est c'est un cas qui qui est très bien pour apprendre parce que c'est en pratique tous les modèles de R se trouvent dans un labyrinthe. Un labyrinthe qui est en dimension beaucoup plus grande qu'un labyrinthe de D potentiellement dans un espace continu mais en dans en pratique on a des États et on se déplace dans notre espace des États avec des États qui sont impossibles. Donc on peut pas aller tout droit jusqu'à notre État final et à chaque à chaque nouvel État. Il faut qu'on choisisse dans quelle direction on veut. Ok, et il y a une petite difficulté qui est que. Qui est que parfois on peut-on peut essayer d'aller dans un dans un État mais pas y arriver. Ça on on va, on on va en parler tout à l'heure. Donc pour l'instant on a l'image du l'image du labyrinthe et donc la poésie c'est le modèle qui dit à partir de mon État, quelle action je fais. Donc dans le cas du labyrinthe, l'action c'est au bas droite gauche. Et et mais ça, on peut imaginer ça n'importe n'importe quel modèle. Ensuite on va commencer à faire des des probas j'imagine. Niveau probas statistique, vous êtes en statistique, c'est pas c'est pas très grave, mais probas vous êtes tous OK, vous pouvez manipuler, c'est normal. Donc la politique elle peut être déterministe, c'est à dire OK, si je suis dans cet état là, je fais ça. Elle peut aussi être stochastique, c'est-à-dire que le modèle au lieu d'apprendre. Au lieu de comment dire, au lieu de faire toujours le même choix dans la même situation, il va pouvoir avoir des probabilités, c'est à dire c'est pas moi. Si, si je suis dans le labyrinthe. Si je suis ici par exemple. On peut imaginer en fait que ici, il y a 2 chemins qui sont possibles, qui sont équivalents. Bah en fait on peut imaginer que le la politique optimale ce soit au lieu que ce soit toujours aller en haut. Bah 50% de chances d'aller en gauche, d'aller en haut, 50% de chances d'aller en à droite. Et donc là on c'est on appelle une politique stochastique donc la différence en termes de notation vous allez voir c'est juste. Soit on AA égal PI de S, soit donc là dans ce cas on a PI qui est vraiment une fonction et dans le cas STOCHASTIQUE PI ça devient une probabilité et donc on parle pas. On dit plus a égal PI de s, on dit que on dit qu'on fera l'action a avec la probabilité PI 2A sachant s voilà ce que c'est clair. Tout le monde. Ça c'est important de bien l'avoir en tête parce que vous allez voir on a il y a beaucoup de mélanges de notation un peu partout et c'est très important de savoir toujours ce qu'on manipule et ce que notre Monty c'est juste une fonction bête déterministe. Est ce que c'est une probabilité ? Y aura y a beaucoup beaucoup de de raccourcis de notation. Un peu partout donc c'est important de bien faire la différence. Donc sachez que les 2 sont possibles et en fonction de ce que vous allez voir, si vous voyez un pyjama à h 1S donc juste un pied s, Eh Ben c'est pas la même chose. Euh Ouais, des questions-là dessus oui. On parle de un. Ah, c'est l'action. Donc dans le cas du labyrinthe, c'est o ou enfin au bas de gauche. Ouais, une question. Ok, super. Euh. Il me semble, je ne je ne garantis pas que dans le cours quand c'est un petit a c'est une action précise, quand c'est un grand a c'est la variable aléatoire associé. Ok donc en pratique, PI de a sachant s c'est rigoureusement écrit PI de probabilité de grand a égal petit a sachant notre état s à t égal petit s. C'est bon pour tout le monde. Ensuite, autre définition très importante, la la fonction valeur. Ouais. Donc la fonction valeur, qu'est-ce que c'est ? C'est donc rigoureusement. C'est l'espérance de l'ensemble des récompenses qu'on va obtenir sur notre sur notre épisode épisode aussi à noter ce que je vais l'utiliser pas mal. Un épisode c'est le. Le fait de de partir de l'état initial à l'État terminal, c'est-à-dire pour je sais pas, pour un pour labyrinthe, un épisode, ce sera le moment. Enfin toute la séquence entre le moment où je je lâche mon agent dans le labyrinthe et le moment où il arrive où qui se perd. Donc un épisode c'est un un jeu quoi. Ok donc on parle pareil, on parle d'épisode et de. Batch d'épisode quand on fait de l'entraînement directement donc la fonction de valeur c'est l'espérance de toutes nos récompenses. Sachant l'état, on se trouve maintenant. Donc le truc important c'est que la la fonction valeur value fonction elle elle elle prend en entrée s. Ok, donc ça veut dire donc dans le cas du labyrinthe ? La fonction valeur c'est si je me trouve à cet État là, à quel point c'est bien d'être à cet État là. Ok, et ça va me permettre en fait d'évoluer dans notre, dans notre. De notre environnement, parce que si on sait que l'État d'à côté, il est mieux que l'État dans lequel on se trouve maintenant en termes d'évaluation. Bah en fait on a tout intérêt à choisir l'état d'à côté et pas l'autre qui a un qui a une value, enfin qui a une fonction valeur qui est qui est plus petite. Ok, donc très très important ça je voudrais la définition exacte tout à l'heure. En un petit détail. Qui est que ? Avoir une récompense à l'instant T c'est pas pareil que d'avoir une récompense à à l'instant T +1OK si on si on est un être humain c'est important de de manger maintenant aussi c'est bien d'avoir son diplôme à la fin dans un an et si vous mangez pas maintenant en fait vous allez pas pouvoir arriver à votre objectif. Donc c'est pas pareil d'avoir une résultat un. Une récompense très grande plus tard et d'avoir une une récompense tout de suite. Donc en pratique pour que le modèle il se concentre quand même un petit peu sur l'étape qu'il y a à faire vraiment à T +1 on a un ce gamma là qui déforme Factor. Qui va venir ? Qui va venir pondérer l'impact de chaque récompense ? Ok, en pratique, quand on est à l'État athée, on connaît pas encore toute la récompense. Donc ce qui est fondamental c'est que la fonction valeur là en fait c'est un peu le gral parce que si on l'a pour tous nos États en fait c'est super simple, il suffit de d'aller vers les États ou la fonction valeur est la plus grande. Parce que comme on sait que l'espérance de la récompense est la plus grande, en fait il suffit de rester sur ces États là et tout se passe bien. Sauf que le problème c'est qu'en pratique cette fonction là on l'a pas bien entendu et donc ce qu'on fait c'est que on ne peut que l'estimer. Et donc vous allez voir toutes les méthodes qu'on va faire plus tard, ça va être en fait en pratique ce sera estimé cette fameuse fonction valeur. Il y a une autre que dont je vais vous parler juste après mais mais celle-là est vraiment, c'est vraiment la fonction centrale. Oui vous pouvez expliquer par rapport aux États. Alors donc cette fonction, elle prend en entrée l'État dans lequel on se trouve maintenant. Donc c'est l'espérance de la somme des récompenses pondérées. Qui en ont aussi que je vous dessiner tout à l'heure. C'est l'espérance de cette somme là sachant que je suis maintenant dans dans l'État s donc typiquement vous êtes aux échecs. Bah l'état s ou vous êtes à un un coût de l'échec et MAT il a pas une fonction de valeur très grande parce que vous êtes à peu près sûr vous allez perdre. Tu vois ? Donc ça veut dire que je sais que dans cet État là en fait l'espérance de mon gain. Pardon le gain en fait. Bon j'ai lâché le mot mais le gain c'est ça, c'est la fameuse somme des la fameuse somme des récompenses. Donc l'espérance de de toutes mes récompenses. En fait elle est très faible parce que je sais que j'ai probablement perdre. Donc en fait je je suis pas dans une bonne situation. C'est ça qui est important, la fonction valeur, la sa, la fonction valeur. Elle est là pour quantifier à quel point c'est bien de se trouver dans un État dans l'État partie. Ouais, parce que à chaque fois que on on est dans dans un cas de processus d'édition, Marco vient. Comme on a vu tout à l'heure, à chaque pas de temps. On reçoit, on reçoit une observation. On fait une action et on reçoit une récompense. Ok, donc à chaque pas de temps j'ai un nouvel État. Non, pas suivant. Ouais, on là on est athée. Donc on va recevoir les récompenses et ensuite on va, on va faire une nouvelle action, recevoir une nouvelle récompense. C'est que la, pour tout le monde, c'est très important. Franchement, n'hésitez pas si y a un truc qui manque. Gamma c'est le cdiscount Factor fact, je sais même pas comment traduire ça. Et il sert à pondérer, on va. Celle qui est fait pour ça, il sert à pondérer les États suivants, c'est à dire qu'on va quand même chercher. Là on est d'accord que la value function elle prend en entrée. Enfin pardon c'est l'espérance de toutes les récompenses sur tout le sur tout le jeu. Et donc si vous avez un jeu où il est quand même important de de de s'en sortir à l'instant T et pas uniquement de gagner à l'instant à l'instant terminal, bah pour faire en sorte que le modèle il se concentre un petit peu quand même sur les États juste après. Et Ben on on va, on va réduire la le poids. De des derniers États avec ce fameux cdiscount Factor. Et il a des propriétés. Donc le fait qu'on l'écrive comme ça c'est pas anodin ce qu'on pourrait dire. Je sais pas, on aurait pu le dire, gamma 1/2, gamma 1/4 de gamma, et cetera. Mais le fait qu'on l'utilise sous une forme d'une d'une suite géométrique comme ça, ça a un intérêt. Vous allez voir juste après. Donc dans le cas typiquement si on met gamma égal 0 le modèle il va il va optimiser uniquement sa récompense à l'instant T +1 c'est à dire qu'il s'en fout de gagner à la fin, il veut juste avoir un Max de points tout de suite. Et si il y a égal un ce qui est aussi possible et Ben là le modèle il va optimiser la somme complète donc il peut complètement sacrifier l'État t +1 pour avoir plus de récompenses après. OK mais bon après c'est c'est. C'est typiquement un des facteurs que vous allez un peu enfin tuner dans vos dans vos entraînements, des des valeurs classiques c'est vraiment 99 quatre-vingt-dix-huit. Enfin 0429042010. Pour faire en sorte de d'avoir typiquement une récompense qui est bien répartie sur les disons les 100 prochaines Itérations de votre modèle, mais qui va pas trop trop regarder les 100 dernières parce qu'en fait il en reste 1000 entre les 2 et il faut-il vaut mieux d'abord se concentrer sur les. Est-ce que c'est clair pour le discontacteur et. Ok. Ensuite, euh. Nouvelle notion, la probabilité de transition. Dans le cas du dans le cas du pardon du labyrinthe, ça on va pas utiliser ces choses là, mais c'est quand même une notion importante parce que. Ce que je vous expliquais, c'est que On a beau enfin dans la vraie vie, On a beau vouloir faire une action, il peut se passer plein de choses. Quoi. Je peux vouloir faire l'action, aller jusqu'ici, mais ça se trouve le sol, il va se casser sous mes pieds et je vais tomber. Donc il y a une probabilité pour que même si l'action que je fais enfin, même si je fais l'action dans le but d'aller dans un nouvel État, il y a une probabilité pour que en fait. Je n'arrive pas à ce nouvel État parce que y a plein de facteurs extérieurs, et cetera, et donc on rajoute une notion de probabilité de transition. Qui est. Quelle est la probabilité d'arriver à l'État s prime à t +1 sachant que je suis dans l'état s à T et que je fais l'action AOK ? Donc mathématiquement ça s'écrit comme ça mais l'idée c'est OK, j'ai une action, j'ai un État. Quelle est la probabilité que j'arrive à l'état du d'après ou qui a la probabilité que une météorite me tombe dessus et je me. Ok donc ça il faut garder à l'esprit on va pas toujours utiliser une typiquement dans le cas du labyrinthe si vous êtes à la case. Si vous êtes à la case du bas et que vous allez à droite, la probabilité d'arriver à droite Elle est. Donc on va pas écrire ces probabilités de transition. Ok, cependant elles existent et parfois on va les, on va les utiliser donc faut les garder en tête. Faudra pas s'étonner si dans les dans l'écriture d'une espérance, et cetera, on on tombe sur une fameuse. Sur une fameuse priorité de de transition. Et j'avais oublié un truc au-dessus. Non, c'est bon. Du coup, quand on a une probabilité de transition, la reward devient aussi stochastique. Donc la reward ? La récompense ça va être l'espérance de notre reward à t +1 sachant que on est à STOK parce que la fonction reward est une fonction qui dépend. Qui dépend de St +1 donc si on n'est pas sûr d'arriver à rester +1 bah en fait on n'est pas sûr d'avoir la reward associé à 1+1OK. Donc là encore une fois, vraiment, c'est important de bien garder toujours à l'esprit ouais. Ouais. Très important de toujours garder à l'esprit ce qu'on manipule est ce que mon R c'est stochastique est ce que mon mon PI est stochastique est-ce que mon État est stochastique ? Voilà parce que on va toujours écrire des espérances, on va manipuler des choses et donc si vous savez pas ce que vous n'aimez, vous allez pas comprendre les passages que je vais faire quand on va faire quelques calculs et là on va avoir des problèmes, OK ? Ok, donc l'État, donc l'exemple du du labyrinthe. Donc je vous ai déjà pas mal expliqué, le seul petit détail c'est que maintenant ? À votre avis, quelles récompenses on va mettre pour le labyrinthe ? On sait vous savez quel État on utilise. Vous savez quelle action on peut faire dans labyrinthe maintenant, moi, moi qui programme mon modèle, qui va choisir de de se déplacer dans le labyrinthe, ce que je choisis comme récompense, ouais. Ouais mais du coup c'est moi qui la programme donc je je connais l'État. Quelles récompenses ? Je vais dire au mot, je vais donner au modèle. Ouais. Ouais exactement. Donc plus le modèle va être proche va arriver, pour plus on va le dire, c'est bien. Non, pardon, pardon, pardon, j'ai rien dit. En pratique, on va faire encore mieux. On peut faire encore plus simple que ça. On va juste lui donner une récompense de un si il arrive et de moins un par. Par temps passé dans la viande à dire on va dire, c'est pas bien tant qu'il est pas arrivé. Ok. Donc ça veut dire que le modèle va devoir aller le plus vite possible à l'arrivée. Et justement ce cette notion de plus plus on se rapproche mieux c'est. En fait ça va être la value function. Bon c'est pas c'est pas très grave mais intuitivement ce qu'on ce qu'on se dit c'est que si on a une grande récompense à l'arrivée et qu'on a moins un pour chaque pas qu'on fait dans le labyrinthe la value function qui est l'espérance de mon gain. En fait, ça va être. Ben en fait plus je suis proche plus mon espérance va être grande. Parce que si je suis à 2 pas de l'arrivée, bah en fait je sais que j'aurais 2 fois moins un et une fois +10 parce que je vais arriver. Donc mon espérance va être grande. Donc ma value function qui est l'espérance de de ma récompense sera grand oui. Ouais mais justement on sera pas très proche en termes de en termes de distance vraie donc donc la value function typiquement ici. 123 4567891011, 2. Tu vois en comptant tout le chemin. Mais. Remarque, remarque très pertinente, oui, y a une question au fond je crois. Non, c'est bon. Ok donc ça typiquement c'est le c'est le TP de demain donc vous allez pouvoir observer ça concrètement. Donc les États, les actions, la récompense, c'est clair. Et maintenant la politique, donc la police du modèle, qu'est ce que c'est ? C'est pour chaque État vers où je vais, donc en pratique on peut le représenter comme ça, avec une petite flèche pour chaque État, dans quelle direction ce que je veux. Donc là, typiquement, c'est l'exemple de la la politique optimale, en fait la seule, la seule qui est possible pour se sortir déjà bien. Pour chaque État, je vais dans la bonne direction. C'est clair pour tout le monde. Oui. Un environnement plus complet ? Complètement tout le but et même même là en soit tu la connais pas, moi je te donne le le le le labyrinthe bien entendu. Là forcément vous voyez tout en même temps vous savez dans quelle direction on va, mais faut s'imaginer que l'agent il est à l'intérieur du labyrinthe donc il sait pas que ensuite se tourner à droite ou à gauche, tout ce qu'il sait c'est que à l'instant T il se retrouve dans un État. Et il faut qu'il choisisse dans quelle direction aller. Et à la fin, donc le nous, ce qu'on nous, ce qu'on aimerait, c'est que le modèle tout seul il apprenne ça. Ok et donc là on est typiquement dans dans le cas de la de la fin du function. Si on a une espérance. On a dit qu'on avait un pour, dans ce cas-là c'est même 0 donc on a une récompense de 0 pour pour arriver juste, on met moins un à chaque fois qu'on qu'on fait une étape dans le dans le labyrinthe. Donc en pratique avec la il faut bien, il faut bien noter que ici la value function, elle est indexée par PI. Ça veut dire que sous l'hypothèse que notre politique c'est la meilleure. OK. Ou alors la value function c'est ça ? Ça veut dire quoi ? Ça veut dire que si je suis dans le cas où je fais toujours la bonne décision pour sortir du labyrinthe, alors l'espérance de mon gain. Donc là on a un un gamma qui vaut qui vaut un parce qu'on fait la somme à chaque fois. Ici, ça vaut moins un ce que ça veut dire que dans un j'ai plus que une étape dans le labyrinthe, donc j'aurais plus que moins un en récompense et rien de plus. Ici, c'est moins 2 parce que j'ai 2 étapes avant de terminer, et cetera. Donc on se retrouve bien avec une value function qui nous donne, qui nous donne l'information qu'on veut parce qu'en fait si on a ça maintenant pour le modèle c'est super simple, il a juste aller dans la direction ou la value function est la plus grande ? Ouais. Je pense que j'en ai perdu quelques-uns. Là non, on l'a pas. Par contre, c'est une première étape. Pour se dire, parce que maintenant, si j'ai ça, si j'ai l'espérance à chaque fois, alors j'en déduis la politique. En supposant donc là en l'occurrence, c'est plutôt dans le sens inverse. Si on suppose qu'on a la politique, alors on a cette cette fonction valeur. Et encore une fois, tout est très lié. Parfois on va vouloir vous allez voir dans les algorithmes, on va implémenter demain. Parfois on va d'abord réduire la la value function et on en on en extraira la politique là dans ce cas-là si on a la value function alors c'est très simple de faire la politique. On a juste à aller dans la direction de faire une montée de gradient sur la value function. Si on a la politique optimale on en déduit la value function aussi parce que à chaque fois qu'on a vu qu'on a la politique optimale, bah on sait que. Si on a l'état, alors on va aller à. On sait quelle action va faire, donc on sait qu'elle récompense, on va avoir et cetera, et on peut calculer une telle fonction. Et donc tout ça c'est très ça va ensemble quoi. Si on a l'un, on a l'autre et on va choisir de calculer l'un plutôt que l'autre ? Oui. Connaître tous les États. Bah le là oui si on la connaît pas justement après tout le problème ça de la calculer, mais là ici on est dans le cas où on connaît. Enfin on vient de on vient de résoudre notre problème, on a ça et je vous dis vous avez la politique, vous avez PI, il est là, je le définis comme ça. On sait pas comment on l'obtient pour l'instant. Si vous avez pis. Maintenant, on peut calculer la value fonction. Vous êtes impossible parce qu'on connaît tous les. Du coup ? Oui, là on est en état liste complète. Ou en probabiliste avec une probabilité de un mais bref, ouais, c'est très bien. Très bonne remarque. Ok, quelques notions à à noter, on va revenir dessus plus tard et donc maintenant vous avez un peu les. Maintenant que vous avez un peu les les premières notions de c'est quoi value ? C'est quoi reward ? Et cetera ? C'est quoi policy ? Il y a des des des notions en RL qui permettent de décrire la méthode qu'on est en train d'employer pour résoudre le problème. Donc typiquement, si je vous dis qu'on va faire une méthode value base, ça veut dire qu'on va typiquement pas chercher à déterminer la porte ici et on va directement essayer de déterminer V et à partir de V on sait qu'on peut en extraire la politique optimale. Ok donc ça c'est value base, on n'a pas de policiers, donc typiquement notre politique à la fin ce sera juste là j'ai ma valise. Je veux juste dans la direction où m'a value fonctionne la plus grande. Donc donc value value base pour les sybase. Acteur critique, ça sera pour plus tard, mais acteur critique, on a les 2. Donc c'est dans les cas plus compliqués où parce que vous vous rendez compte qu'on est pas toujours dans un labyrinthe. On connaît tous les États dans la vraie vie, chaque État que je que je. Que je rencontre est un nouvel État, enfin j'ai je j'ai jamais été allé fraye, là je suis un nouvel État pour moi, j'ai je me suis-je suis jamais venu. Il dit ici je me suis jamais dit Ah mais en fait il faut que je fasse tel pas à droite ou à gauche pour avoir la meilleure relation. Donc en pratique dans la vie on peut pas tout savoir et donc on va on va, on va tout modéliser parce que du coup on aura besoin de temps en temps donc on aura à la fois la value function et à la fois la politique. Et dans ce cas-là, on appelle ça acteur critique. Parce que l'acteur c'est qui ? C'est celui qui détermine la politique et le critique, c'est celui qui dit c'est cet État là, il est bien cet État là, il est pas bien. Et donc, typiquement, qu'est-ce que, qu'est-ce que, qu'est-ce que calcule un modèle qui dit cet État là il est bien cet état là il est pas bien ? Qu'est-ce qu'est-ce que estime le critique dans ce cas-là ? Je vous ai dit le critique, il dit, cet État là, il est bien ou cet état là il est pas bien ? Ça veut dire comment ? Non, pas la vie World. Il n'y a plus que 3, c'est pas. La question c'est dans le modèle acteur critique on a on évalue les 2, c'est-à-dire que dans le cas du labyrinthe on a vu que il suffisait d'avoir la value function, mais dans le cas dans des cas plus compliqués, on ne peut pas avoir accès aux 2 parce que en fait c'est des États, continents, et cetera. Euh bah il faut qu'on évalue les 2, on va un, on va avoir un modèle de politique et un modèle de critique, donc qui va évaluer ce qu'on est en train de faire. Donc le critique il va dire à chaque instant ça, ce que tu fais, c'est bien ça ce que tu fais c'est pas bien. Donc quand un critique en à partir de l'État. Donne une valeur pour dire si le modèle fait quelque chose de bien ou pas ce qu'il évalue. La value function. Ok. Parce que la value function c'est quoi ? C'est la fonction qui évalue à partir de l'État s si enfin l'espérance de notre de notre devoir à la fin. Ok, tu prends des sourcils, t'as pas loin ? Ouais c'est bon, tu prends ça les sourcils ? Très bien ça, non mais c'est pas ça, c'était pas une. Ok bon ensuite donc des notions à on y reviendra mais c'est des notions que vous allez rencontrer. Donc c'est bien de les avoir en tête. Et ensuite modèle Free modèle base ça veut dire que on va modéliser ou non l'environnement. Euh. On reviendra dessus c'est ça hein, de de ces orage ? Écoutez, il est 9 h, on peut faire une pause maintenant ce qu'ont des questions bah hésitez pas et et ensuite on reprendra les définitions vraiment mathématiques. Je sais pas alors attends. On va passer maintenant aux définitions plus mathématiques de qu'on va utiliser en RL et donc comme je disais tout à l'heure on va, on va parler de processus de décision, markovien ou MDP ? Et là là, on va rentrer dans des définitions qu'il faut vraiment bien connaître et on va essayer de de voir un peu un peu les maths derrière. Alors est-ce que vous savez ce que c'est une chaîne de Markov ? Parce que ça vous parle pas du tout. Ouais Ouais, vous connaissez de nom quoi ? Bah une chaîne Markov c'est super simple. Euh vraiment c'est une scène de Marco, c'est. C'est le cadre mathématique pour décrire une suite de variables aléatoires, mais une suite un peu particulière. Donc si je si je dessine ici mon État St. C'est +1 ? Et ce grand t dernier État final, ça c'est une suite de variables aléatoires OK et une chaîne de markoff donc ça cette suite de variables natatoires là elle devient une chaîne de Markov. Si l'État St +1 dépend uniquement de St. Ok. C'est une hypothèse très très forte qui dit que mon État à l'instant T +1 il ne dépend que de St et de rien d'autre c'est à dire si je fais une bêtise au début à l'instant s un à l'instant 5. En fait on a complètement oublié. On se souvient que de s 4. Ok, donc ça c'est. Une chaîne de Markov, OK. Donc mathématiquement, ça veut dire qu'on n'a pas de dépendance. Donc markoff. Donc ça, c'est le. Mathématiquement, ça veut dire que la probabilité de l'État est St +1. Sachant. Mon est à un et cetera, mon État St. Et ça, c'est égal. À la probabilité. De St +1. Sachant rester ou court. Ok, ça c'est la définition de donc c'est la. C'est la définition de la propriété de Markov, vanité de la chaîne. Donc tout ça pour dire. Compliqué pour juste dire que l'État t'es +1 dépend que de thé et c'est tout OK et donc ? Le fait de faire cette hypothèse qui est très forte, je le rappelle hein, parce que dans la vraie vie c'est c'est pas du tout vrai. Si vous faites une bêtise à l'instant T, Ben vous avez la jambe cachée, vous vous en rappelez longtemps. Donc ça c'est une une hypothèse qu'on fait pour simplifier l'étude du au début et donc on va s'en libérer plus tard, mais on. En gros, avant le RL, quand on quand on allait pas faire de Deep learning, on était obligé, on avait on avait plus que les maths d'une certaine manière, donc on est obligé. On est obligé de passer par ce genre de choses parce que en fait, quand on commence à à prendre en compte trop d'États ensemble, on s'en sort plus. Mathématiquement c'est c'est pas gérable, donc on se plaçait dans le cadre des chaînes de Markov et on avait plein de de de choses pour. Pour. Pour, comment dire, libérer un peu les contraintes. Mais c'est vraiment le renforcement learning. Il tient ses racines de ça, vraiment. Des modèles qui ne dépendent que de l'instant précédent. Donc toc, toc, toc. Ouais donc formellement on le note donc un processus de markoff maintenant maintenant qu'on sait ce que c'est qu'une chaîne de Markov, on a un processus de Markov sachant que l'objectif là donc on va arriver au Markov décision process. Donc là on a juste on sait ce que c'est markoff maintenant on va regarder c'est quoi un processus de Markov ? Et donc le processus de markoff, c'est les États possibles. Donc tous mes s ? Avec les probabilités de transition. Ok, donc maintenant ? Donc ça on a maintenant on a markoff process. Qu'on définit comme. L'ensemble des. Des États et des probabilités de transition. OK, donc ça, ça vaut quoi, ça vaut ? Tous mes petits tests possibles. Un. Je s trentaine, OK. Parce que c'est clair et donc on sait que ici on notre État, donc notre ensemble de sites, notre suite de variables aléatoires. Il est à valeur. Qui est à valeur dans. Grand-e-s à la puissance. C'est bon pour tout le monde, donc ça c'est notre notre grand test calligraphie est super bien calligraphié. D'ailleurs c'est l'ensemble des États possibles et les les petits tests et les États possibles concrets. C'est-à-dire ça va être les coordonnées, donc 0 0, ça va être un petit test par exemple ça une petit test 0 par exemple. Et grand t, c'est la variable aléatoire qui a été associe la le l'État dans lequel je me trouve. Ok, y a une question. Et puis pardon PA une probabilité c'est l'ensemble. Enfin c'est la la fonction qui va de s 3S calligraphié. Action. 3. Ah. Action et qui associe. On prend enfin. Je voulais faire, c'est plus simple la probabilité. De passer à esprit sachant. Une petite esprit, sachant que je suis dans l'État et que je fais l'action a ? Ça c'est les probabilités des transitions. Pardon, il y a pas de y a pas d'action pour l'instant. Justement quand on va rajouter l'action c'est pour passer à 1MDP, donc là on a un ensemble, on a une variable aléatoire qui se balade dans grand est. Ok, et on a des probabilités, zut. Hop. Euh, on a des probabilités de transition ? Ok, donc. On va avoir y a. Durant une vision tout à l'heure. Mais donc on est dans les dans l'État 0 on est dans la case tout en bas à gauche du labyrinthe. C'est quoi la probabilité de se retrouver dans la case de droite à côté ? La case de gauche à côté ? OK donc la probabilité c'est notre épée et notre c'est l'ensemble des États possibles. OK donc là pour l'instant. C'est pas de une force mécano m'a juste, on a juste un un agent. Enfin on a juste un une variateur qui se balade dans un ensemble d'États. Il y a une probabilité de passer d'un État à l'autre. OK on prend aucune décision. Il y a pas de une Ford Mustang je pense. Ouais, une question. Le grand. Je vais gratter, c'est l'ensemble des États prochains possibles, des étapes possibles tout court. De tous les États ? Ouais, exactement. Typiquement, euh. Ici dans le cas de. Ou. Dans le cas, dans le cas de. Du labyrinthe, c'est égal à ça vous parle sa notation ? J'en sais rien, moi, la longueur. 0L. Cette notation, c'est un intervalle numérique, ça veut dire c'est de 012345 jusqu'à jusqu'à l ? D'ailleurs, je vais plutôt. W. Hop, 0H. Donc ça, c'est. C'est les coordonnées. Mes États, c'est mes coordonnées par contre. Donc j'ai donc la probabilité de passer de tout en bas à gauche à tout en bas à droite. Elle va être nulle du tout, mais elle existe quand même notre notre réalité. Là elle est bien définie de de l'État 0 0 jusqu'à l'État n enfin l enfin oui. Mais mais en pratique j'aurais juste une probabilité non nulle d'aller de de 0 0 à 0 un et de 0010, c'est bon pour tout le monde. Ok, donc maintenant ouais. Ouais. Ponctuel et après on regarde une méta facture. Ouais là ici on a été +1. Ouais, du coup ça veut dire. On a plusieurs, plusieurs étapes passées ou c'est juste ouais alors en gros donc là on est dans un ouais donc on s'intéresse un d'abord la première étape c'est l'ensemble de mes étapes. L'ensemble de tes États c'est une suite de ventilatoires d'accord, on sait pas à priori quel quelle valeur va prendre s 5 parce qu'on sait pas, on sait pas où est-ce qu'on se trouve à ce moment-là dans le labyrinth. Donc ça, c'est l'ensemble des États. La première hypothèse que je fais, c'est que. Est-ce un donc ma s 5 va dépendre uniquement de là où je me trouve juste avant, c'est-à-dire s 4 ? Et ça, c'est une hypothèse forte, parce que dans la vraie vie, si je me déplace donc dans 5 h, en fait là où je serai dans 5 h, ça va carrément dépendre de là où je suis maintenant. Parce que si là maintenant je suis ici, je serai pas à Miami dans 5 h parce que si on a de la télé mais donc là je fais l'hypothèse que mon État était +1 ne dépend que de l'État AT. Et ça permet donc, et donc mathématiquement ça s'écrit comme ça parce que a priori en fait. Temporellement St +1 il arrive après s un et St, donc il y a à priori une dépendance causale, donc on peut pas dire enfin ça c'est une hypothèse très forte. Mais par contre ce que je dis, c'est que, mais moi je je fais l'hypothèse que. En fait, ça dépend que de ça. Ok, donc super. Maintenant on a un on a un MP toc, toc, toc, c'est bon pour tout le monde. Les probabilités de transition ? Ouais. Stream point S. Si c'est trop vite ? Ouais, c'est ça p de s prime sachant s. C'est la probabilité que mon État a été +1 ce soir exprime sachant que mon État ATT c'est s ? P c'est probabilité. Oui, désolé. Ouais, notre notation. C'est clair du coup ? Ok. Donc maintenant on a un marqueur, un processus markovien, OK ? Maintenant, on va vouloir interagir avec ce processus parce que là, pour l'instant, on n'a pas d'argent dans l'histoire. Le processus il se passe tout seul et et on a on a pas de, on n'a pas de modèle à apprendre. Euh bon voilà du coup ça c'est la propriété de markoff. Ça du coup, c'est clair pour tout le monde, reste bien de décrire plutôt que juste l'arme de montrer au tableau. Maintenant, on va rajouter une récompense. On a un processus markovien et maintenant on va venir juger dans le temps si l'état dans lequel on se trouve il a, il a, il a une récompense ou pas ? Ok donc je rajoute la fonction R. Qui est. Qui est l'espérance de ma reward à l'étage suivant, sachant que je suis dans un État en particulier ? Donc c'est là donc là encore une fois on est dans le cas où on est stochastique. Donc on sait que. Enfin, dans le cadre du labyrinthe, on sait que la reward ce sera moins un dans tous les cas, mais dans le cas général, dans le cas d'un markoff reward process, on. On définit R comme étant l'espérance de ma de la reward que je vais récupérer à l'instant t'épuisant. Donc maintenant on rajoute une nouvelle. On rajoute, on rajoute l'ensemble des rewards. Ok. Bon, du coup, j'aurais dû l'écrire là plutôt. Maintenant, on AMRP. Plus. C'est bon pour tout le monde. Bon MDP, en pratique c'est pas trop utilisé, c'est à dire qu'on a une récompense, on a envie de faire des actions. Donc voilà c'est c'est juste, c'est juste pour faire la transition. Mais donc le mrp on le définit comme SP R et on rajoute le gamma parce que du coup on définit la value function pour 1MP, pas pour 1MDP mais n'est pas très important. Donc on peut directement. Oui on va on donc pour pour le MP, on va du coup définir cette fameuse somme géométrique là. Euh. Bon j'ai pas vous l'écrire et déjà elle est déjà, elle est déjà ici, mais la nouvelle définition très importante, le le gain ou la ou le la récompense cumulée cumulative reward. Donc là encore une fois il y a plein de termes. Je suis désolé, c'est ça c'est le c'est l'i a on chaque chacun y rajoute en nouveauté un avec un de vos noms alors que ça ça se faisait depuis des années. Ben donc Notez bien tous les. Toutes les variations cumulatives, reward, gain. Il y a, je crois, que même de te. C'est déjà bien sûr votre, mais ça c'est le grand G et c'est cette fameuse somme pondérée des reward aux États suivants. OK donc ça c'est une définition à connaître est très facile donc enfin apprenez là par cœur. Mon gain raté c'est Marie Whate plus gamma rebat +1 plus gamma carré et cetera avec donc une somme qui est pondérée par Mesdames et associer à ce gain. Je peux maintenant définir rigoureusement ma value function dont je vous parlais tout à l'heure, qui se définit rigoureusement comme. L'espérance. De mon gain sachant mon État athée. C'est bon pour tout le monde. Donc maintenant ma value function. On est d'accord, c'est c'est une fonction, c'est pas une variable. Issoire. Ok. Ici, on a un petit V 2S et on n'a pas un grand VOK. Attention parce que en fonction de ce qu'on définit, ça peut devenir une aléatoire juste dans ce cas-là. La définition dans le cas du du Mark of the Work Process c'est une espérance. Donc c'est une fonction. OK, ça dépend uniquement du petit test qu'on met à l'intérieur dans le cas du labyrinthe. Vous vous avez remarqué que on avait du définir une policier pour ça ? On est d'accord, sauf que là, pour l'instant on n'a pas de décision. Donc on n'a pas besoin de enfin on comment dire, on a pas d'action à prendre pour avoir l'État suivant l'état suivant uniquement défini par notre propriété de transition ici. Ok. Donc là encore une fois, c'est c'est souvent des trucs qui peuvent. On peut prendre la tête, mais bien prenez bien le réflexe de vous dire OK cette fois là mon V c'est une fonction. OK c'est comme ça qu'on le définit. Et on va voir dans le cas où on va rajouter des actions, ça va, ça va changer de statut d'une certaine manière parce qu'on va rajouter de la complexité. Mais dans le cas d'un, d'un, d'un MP ou juste, on n'a pas de décision à prendre. Cette cette fonction là, on la définit comme ça et de manière générale, on la définit toujours comme ça. C'est juste le gain va changer du coup, et cetera. Enfin les les, les dépendances vont changer. Mais la value function, sa définition rigoureuse, c'est ça, c'est l'espérance du gain raté sachant qu'on est dans l'État en particulier. Ok. Hop hop hop. Autre autre équation très importante, l'équation de Bellman. Et là où je vous demande de sortir un papier, un client, vous avez essayé de me montrer ça, donc ça c'est aussi à très bien connaître. C'est une autre manière d'écrire la la fonction de valeur et comme vous avez noté les définitions de la de la value function et du gain, vous allez pouvoir me montrer cette équation là. Voilà, je vous laisse 5 Min. C'est vraiment il y a il y a 3 lignes hein, c'est pas compliqué mais c'est important de de le faire une fois. Pour. Pour comprendre ce qu'on dit. Ouais. De quoi ? Bah oui, ton RT et ton. Bref, ce n'est pas une danse, on s est une chaîne. Bébé. Levé ? C'est pas une chaîne de Markov ça c'est c'est pas une variable, voire le V c'est un petit V donc c'est une fonction. Donc c'est ce truc là ça a une valeur particulière. Donc tu vois comme tu mets une espérance, tu passes d'une valeur natoire, tu passes une variable aléatoire à une une valeur. Donc il y a plus de stochasticité ici. L'espérance elle est elle coupe, elle casse la Stochasticité tu détermine ton truc. Oui c'est une ambulatoire mais ce que du coup là ça devient l'espérance de tout ça. Ce terme là, c'est c'est une armée. Je veux que tu me montres ça. Vous savez que vous savez que la value function s'écrit comme ça. Comme ça là ça c'est la définition de la value function et maintenant à partir de ça. Je veux que tu me montres ça. C'est pas la même chose d'accord tu vois là on a que des rewards à l'intérieur. Maintenant, on fait intervenir la vallée Fashion. Non, c'est pas clair. Hop. Ouais vous voulez vous voulez faire apparaître le V de St +1 ici ? Donc vous passez de la définition ici avec une somme uniquement des laboratoires reliés à la récompense voulez faire apparaître ? La value fonctionnel était plus là. Oui. Hop. Vous pouvez l'écrire comme ça pour éviter d'avoir à gérer les index m'écrivez pour. Vous l'écrivez en gros pour t égal 0 à l'infini. Sachant que s 0 égal qui revient au même hein, comme ça c'est plus simple à écrire. Quelqu'un a réussi. Oui, c'est la, la preuve est très simple, hein ? La preuve est très simple, mais je serai. Je suis curieux de savoir si quelqu'un arrivera à me la justifier un entier. Vous m'avez dit au début du cours que vous vouliez faire des maths un peu. Ne vous inquiétez pas, on en fera pas beaucoup in fine, mais je pense que ça c'est vraiment important, c'est vraiment la base. Quelqu'un aussi. J'ai be be en fait, je peux connaître la value function ici et hop, je propage, je propage, je propage, je propage au coupage et c'est comme ça qu'on arrive à calculer la, la vallée fonctionne sur tout le labyrinthe en entier. Là j'ai 0 comme récompense, j'ai gagné. Parce qu'à partir de ce moment-là, je sais que ma value punition ici c'est moins un parce que c'est y a plus qu'une seule étape. Et à partir du moment où je fais ça, Ben là c'est moins 2, là c'est moins 3 c'est moins ouais. Non, j'en connais un, tu connais St +1 et t'en déduit St. Et du coup en fait tu le fais, tu le fais rétroactivement, ça veut dire t'explores tout tu connais St moins un tu viens découvrir RT tu fais ? Ah mais du coup je sais que à T moins un elle m'a value fonction c'était ça ? Ok. Oui. Si du coup, c'est la jambe qui va explorer tout. Il va tout tester en pratique. Voilà. Du coup on est sur un cas très simple, mais dans l'idée. Pour avoir un modèle qui fait ce genre de choses. Donc on est dans le cadre de du labyrinthe donc on est d'accord c'est un super simple et dans l'idée t'as accès à toute la value function parce que t'as tout exploré et c'est un peu ça le le l'intuition c'est que là on on est sur des problèmes très simples, on peut explorer tout l'espace et donc si on veut connaître toute l'information sur. C'est quoi l'espérance de mon gain si je suis ici ? Il faut bien avoir tout fait avant pour pour pouvoir le savoir. Je sais pas si truc là. Très bien. Ok. Alors où ce qu'on va dire ? Ok. Maintenant. Un petit exemple ? Euh, le Markov Process est un étudiant. Donc il y a plusieurs étapes possibles, y a l'étape Facebook et donc les flèches, ça symbolise la possibilité de passer d'un État à un autre. Donc là quand on est dans l'État Facebook, on a une probabilité de 90% de rester sur Facebook ou une probabilité de 0 un de d'aller en classe. Quand on est en classe, on a une probabilité 0 5 d'aller dans la classe suivante, et cetera, et à la fin. Nous, ce qu'on veut, c'est arriver à l'État. Je, j'ai mon examen. Du coup à chaque État on associe une récompense non ? Typiquement on a moins 2 pour chaque classe parce que les classes c'est super chiant. Et puis on arrive à l'examen et on a +10 si on passe. Euh. Il y a l'étape pub en dessous, après les cours et l'État et l'État, ce qui bien entendu. Donc. Là on a, on reconnaît 1RP, on a des États possibles. OK, on a tous nos États, on a notre fonction p de probabilité qui associe la probabilité de de se trouver dans un État à partir de l'État précédent, donc qu'on représente par des flèches. Donc bien entendu l'État. Enfin la probabilité de passer de Facebook à pass la ligne nulle parce que y a pas de flèche. OK. Souvent on représente des modèles, les modèles markoviens comme ça, c'est pour ça que je vous ai fait un petit graphique comme ça là haut ? Donc on a une chaîne de markoff, on a juste une chaîne comme ça justement pour ça s'appelle une chaîne. Et vous voyez que ici, on a pas une chaîne de Markov en soi. Enfin si, pardon. Du coup les le la suite des États, c'est une scène de Malakoff. Mais la chaîne se déplace dans ce graphe là qui donne tous les États possibles. Ok donc là ça vous dit à partir de l'État X, vers quel État je peux aller, quelle récompense j'ai et c'est quoi la probabilité de d'aller dans l'autre ? Donc. D'abord. L'effet du du cdiscount factor ? Donc là je vous dis pas comment est-ce que on a trouvé exactement la la toute la value function mais on suppose qu'on qu'on l'a ? Et donc ici, on met. Je vous donne directement. La value function avec gamma égal 0 comme ça Vous voyez ce que ça fait. Ok donc gamma 0 ça veut dire quoi ? Ça veut dire que on prend uniquement la reward à l'instant T ? Du coup c'est super simple, là on a moins 2-2+10+1 0-1 en fait c'est exactement ça. Parce qu'on s'en fout de savoir dans combien de temps. On s'en fiche de savoir c'est quoi ma récompense à l'instant d'après, on veut juste la récompense maintenant. Ok, donc ça c'est un étudiant. Un peu un peu dommage pour lui mais il veut juste optimiser son son bonheur tout de suite quoi, OK ? Euh donc là on voit que gamma égal 0 c'est pas top ce que ça veut dire que typiquement si on est si on est dans la classe une, on va préférer aller sur Facebook plutôt que d'aller dans la classe 2 ? Parce que la moins moins un, c'est moins pire que moins 2. Et on peut pas malheureusement pas aller au pub parce que le peuple c'est c'est réservé, que on sait qu'on fait la Castro. Euh. Donc ça c'est pour gamma 0. Maintenant, si je vous mets un gamma plus classique en une force mot learning 0 9. Bah là tout de suite ça change un peu. On voit que donc là pareil, c'est pas évident de trouver ces valeurs là. Parce que. Comme je vous le disais, il faut qu'on puisse tout est lié, donc il faut pouvoir faire le calcul en supposant qu'on sait le faire. Ça, j'ai les résultats qu'on obtient. Donc on voit que quand on applique le le cdiscount Factor, Bah Facebook typiquement ça passe de moins un à moins 8 quasiment. Alors que passer de la classe une à la classe 2 parce que c'est déjà beaucoup mieux. Parce que en fait on prend en compte le fait que si je vais dans la classe 2 et Ben en fait je me rapproche de l'objectif final qui est de passer mon examen et donc disons que la l'espérance de mon gain final il augmente sachant qu'on prend en compte le le futur. Ok. Oui. Est-ce que là le fait qu'on passe dans la classe 2, sa promotion en compte, le fait qu'on esquive entre guillemets le fait qu'on peut aller sur la classe Facebook et donc éviter une grosse perte, ça. Les abonnements ? Logique il y a exactement si si jamais j'avais tracé un une flèche de classe 2 vers Facebook, probablement que la value function de la classe 2 serait plus faible. Parce que j'ai plus de probabilité donc en termes d'espérance si si, on imagine vraiment le modèle comme ça en termes d'espérance si je suis dans la classe 2, on espérance va être plus faible parce que j'ai toujours une chance de retomber dans Facebook et Facebook. C'est vraiment le minimum local dans lequel il faut pas tomber parce que sinon tu restes dedans et tu perds beaucoup de points OK, donc oui ça influera après c'est. C'est ouais, c'est c'est moins intuitif on va dire. C'est. J'essaie de l'expliquer intuitivement mais. Mais c'est moins, c'est moins évident. D'autres questions pour le cdiscount Factor, vous comprenez l'utilité ? Donc ça permet vraiment de se repérer dans l'espace. De se repérer dans l'espace et de prendre en compte le futur, sachant qu'il peut y avoir des minimums locaux de récompenses mais qui sont pas intéressants à long terme. Oui. Enfin le gamin, donc, c'est l'importance de la réponse, enfin de la de la récompense sur l'État, juste suivant enfin et après enfin la récompense qu'on aura les récompenses suivantes. Mélo ? Est-ce que y a un facteur sur lequel on peut agir pour ? Plus ou moins répartir l'impact des récompenses dans le temps, c'est-à-dire sur si on veut repartir par exemple sur les 5 actions prochaines ou alors si on veut que Ben millième action elle est plus importante. En soit, tu peux. Simplement le problème que tu vas voir c'est que si tu fais pas cette définition là, bellman n'est plus vrai. Et donc t'auras plus t'auras plus facilement accès à une propagation de de tes États. Mais en soi, ce que tu enfin. Le gamin est justement déjà un peu pour faire ça quoi, si je trace ? Je trace mon gamma. En fonction de t donc là ce scan le maté quoi si vous avez un un gamma de. Vous avez un gamma de un fait, vous allez avoir le même poids partout. Si vous avez un gamma de 0. T'aura juste le premier État et ensuite entre les 2 on a une suite géométrique décroissante puisque gamma est inférieur à un donc tu vas avoir avec un gamma c'est pas 0 99 tu auras. Donc tu vas prendre un peu en compte l'état 100 mais pas trop par rapport tu vas quand même garder t'auras toujours plus d'importance au gamma T +1 par rapport à gamma T +1000 ça tu choisis pas quoi ? Par contre tu vois c'est à quel point ce que je m'intéresse aux gamma 1000. Ou si je m'intéresse uniquement aux au premier, au premier, enfin au premier. Donc ton gamma, c'est vraiment ça tu à quel point est-ce que tu prends en compte tes ? C'est donc on apprend tous au tous en compte un si tu prends tout pareil, c'est le, c'est le labyrinthe de tout à l'heure. Je vais bien de tout à l'heure on avait un gamma de un donc avec un gamma de un le la value function bah en fait on a plus de de gamma pour savoir hein. C'est juste la somme de récompense ce qui peut avoir du sens. Tu vois dans le cas du labyrinthe où en fait on sait qu'on a un nombre limité de d'opérations. Et que on veut en prendre. On veut prendre en compte toutes ces opérations. D'autant plus qu'il y a la récompense est située uniquement à la fin. Là ça a du sens de prendre un gamma de un. Par contre dans le cas de l'étudiant ici. Bah en fait, on veut que l'étudiant, il puisse pouvoir optimiser son son bonheur, enfin sa récompense petit à petit. On veut pas qu'il le fasse juste pour l'instant T +1 parce qu'on a vu que c'était pas bon du coup on prend un gamma de 0 9 voilà donc moi 0 9 pendant 090 99 c'est les. Des valeurs classiques. 0 99, ça correspondra à la 9. Ouais, par exemple, faut. Il faudrait faire le calcul mais 0 99 à la puissance 1000. Je pense pas que le faire maintenant mais il te reste encore un petit peu de poids quoi. Voilà voilà. Hop donc, petit exercice rapidement. Dans le cas où, on imagine qu'on a réussi miraculeusement à calculer toutes les autres valeurs. Euh. Donc là on est dans le cas avec gamma égal 1OK je vous demande de me calculer. La valeur de la value function. Pour l'état de classe 4, classe 3. Ok. Et donc là vous allez utiliser bellman. Ce que vous savez que V de classe 3. Ça vaut la récompense à la classe 3 que vous connaissez, elle est de moins 2. Plus gamma, qui vaut un fois. Euh la la value function de des autres États ? Enfin de l'espérance de mon nouvel État. Donc là si je vous le réécris. On sait qu'on est à la classe 3, donc là je vous demande. Si j'applique bêtement. Enfin, c'est classe 3. Ça vaut. Donc moins 2. Parce que je le connais maintenant, on les déterminait, on est dans l'état classe 3OK, c'est plus une espérance, plus l'espérance. 2V 2, c'est sous un. Sachant que on est dans la classe 3. Là, normalement, vous devriez pouvoir le calculer très facilement. Donc comment est-ce qu'on calcule ça ? Je vous laisse 2 Min pour le faire. Ça prend 30 secondes. Vous avez les probabilités de transition, vous avez les vous avez les value function à côté. Vous pouvez calculer l'espérance. L'espérance c'est quoi ? C'est la somme des probabilités multipliée par la value function. Ok, on a une réponse ici. 4,30. C'est clair pour tout le monde ce qu'on vient de faire. Vous voulez que je l'explique ? Ouais. Alors est-ce que ce passage là il est clair ? Ouais, non. Ok. On part de ça, OK ? On vient de le montrer, on a la preuve, on sait que ça c'est vrai. On m'a valu fonction de l'espace, donc de de mon état, n'importe quel État. Donc là on veut l'État classe 3OK, c'est égal à l'espérance de la récompense à mon État St Ben mon pardon, la récompense de que j'obtiens à. À cet état là. Donc la récompense c'est quoi ? Parce que vous l'avez écrit ? Hop, la récompense je la connais. C'est combien la récompense quand on est ici ? Donc on est d'accord, je peux remplacer mon RP par moins 2 que je le connais donc moins 2. Ok, je le sens de l'espérance parce que c'est une valloire, OK, donc il me reste. L'espérance. 2 la vallée location était +1 donc là on applique la définition de l'espérance que sa définition l'espérance probabilité par exemple Ouais donc moins 2 plus la somme. Pour tous mes États possibles. Donc dans dans tous mes États possibles. Labilité. De passer de de ? La probabilité d'arriver dans petit s exprimer disons, sachant que je suis dans classe 3. Multiplié par. Donc là, la définition de l'espérance, la somme, la probabilité multipliée par la par, non par. Pas V de V de quoi là j'ai indexé par exprime en bas V 2 exprime c'est clairement non, pas ça. Ça, c'est la définition de de l'espérance. Je souhaite tous capables de faire ça, hein ? J'applique la définition de l'espérance ici. Je sais que je suis en classe 3 donc la donc je je fais l'espérance donc l'espérance de m'avoir aléatoire est +1 c'est la somme sur toutes les possibilités de la probabilité d'arriver dans cet État là multiplier par la valeur que je peux prendre. C'est clair ? Donc là en pratique, maintenant qu'on a ça, c'est quoi tous les États possibles ? Bah y a 3 États possibles. Qu'on regarde les flux y a pas y a que 2 États possibles. On est en classe 3, on peut soit aller au pub, soit passer l'examen donc. Ma fonction ici là elle va valoir donc moins 2 plus la probabilité de passer à l'examen 0 6 multipliée par la récompense que j'obtiens à ce moment-là pardon la pardon la value function à cet endroit là que je connais. Donc c'est 10. 0 6\*10. En plus la probabilité d'aller au pub 0 4 multipliée par la value fonction du peuple. C'est bon ou pas ce que c'est clair maintenant pour tout le monde ? Oui, OK. Comment ? C'est clair ou pas ? C'est bon. Moi j'avais utilisé du coup le R. Ouais mais c'est pas R, c'est le V parce que on a l'espérance de V, l'espérance de. Et c'est ça qui serait mondial, c'est que là on veut calculer value function, on veut pas calculer le reward, on la connait. Tout l'intérêt c'est que on est dans le cas où notre étudiant il est perdu, il sait pas dans quelle direction aller, il veut maximiser sa reward. Eh bah Eh bah ça arrive hein. C'est à nous de lui dire Bah en fait moi je l'ai calculé l'espérance de ta reward en fonction de tous tes États donc c'est super simple. T'as juste à aller dans dans l'État qui me donne la meilleure espérance de de reward facile non ? Et donc donc c'est c'est pas évident de calculer cette value function mais là on le fait du coup parce que on a une équation de bellman qui me dit que en gros si je connais tous les États ? Qui suivent potentiellement mon mon État, je peux propager leur valeur vers mon nouvel État. Ok et c'est ce qu'on vient de faire, c'est à dire que dans l'hypothèse où on a calculé toutes les value function pour tous les États possibles. Alors je suis capable de calculer ma nouvelle value function. Pour mon état ? Ok, et ça, ça rend final. On est dans un dans un espace de possibilité. Ce que vous dit l'équation de Delman, c'est que. Si je connais tous les États possibles ? Qui peuvent succéder mon État alors je connais ma value function OK ? Donc ça c'est très important et c'est la base de ce qu'on va faire demain. Encore une fois où on est dans le cas du labyrinthe. Donc le labyrinthe c'est vraiment exactement comme comme notre enfin ce qu'on vient de de faire ici. Juste que c'est un peu plus simple, y a pas de probabilité de de. Il y a il y a pas de boucle de rétroaction et cetera, mais le le principe elle-même. Donc ça veut dire que une fois que j'ai le bout du labyrinthe, je suis capable de rétro propager l'évaluation partout dans le labyrinthe. Et en pratique, même si vous avez la viande très compliqué. Bah grâce à cette méthode là, vous allez pouvoir retrouver votre chemin dans le labyrinthe de manière automatique. Ce qui est cool déjà c'est un c'est un, c'est une première étape. On n'est pas encore à faire bouger des les robots de penhouët, mais on n'est pas loin. Ok. Maintenant, le Boss final. On a un Mark reward process. Maintenant, on va passer aux fameux Markov décision process, donc on va rajouter le fait que l'agent peut prendre une décision. Là avant notre étudiant, là il avait pas le choix, il avait une probabilité. C'est un peu une étude globale, une étude globale sur ce que font les étudiants. Là maintenant on va, on va prendre des décisions. Donc dans notre MRP, on avait la probabilité. Les États ? La remorque et là on va ajouter le a qui est quoi, c'est quoi le a l'action, les actions, la possibilité qu'a le modèle de prendre une action dans l'environnement. Ok donc le MDP à bien connaître, on a État action les probabilités de transition et et les actions les. Tac TAC TAC, on peut le dessiner comme ça aussi. Et donc maintenant on a en vert ici, ça va être très clair pour ceux qui sont au fond en vert, c'est les États et on rajoute les petites petites boules rouges ici qui sont les actions. C'est à dire que quand je suis dans l'état classe une, je peux. Prendre la décision d'aller au pub ? Mais il y a une probabilité pour laquelle je me passera très bien par mon prof qui va va m'emmener dans la classe 2OK. Donc on a plus. On peut prendre une décision, mais on n'a pas non plus le contrôle total, OK ? Donc donc ce cas-là c'est vraiment le cas complet du Marc de son process. En pratique, on verra que souvent on aura des actions qui seront déterministes, c'est-à-dire que dans le labyrinthe, si je prends la décision d'aller à droite, personne va m'en empêcher. Donc en pratique on va couper là le petit 0 5 ce sera juste 111 quoi on va dire s un en prenant l'action un on va rester dans le un. Ce qui veut dire c'est bon. En attendant, on est sur le cas général où on peut prendre une décision d'action, mais il on peut se retrouver dans dans l'État, dans un autre État auquel on s'attendait. Ok. Ouais, y a juste d'autres exemples. Pour les actions, ce que j'en ai juste après, non. Par exemple, vous voyez les machines à sous là où vous pouvez choisir un un bras attiré là. Eh Ben c'est un peu ça. Vous avez 3 machines à sous. Et vous savez pas quelle est la probabilité de gagner sur chacune des autres ? Y en a une qui est pétée et genre la probabilité de gagner c'est 0 9. Vous avez tout intérêt à en fait tirer que sur celle-là quoi. Et Ben du coup donc là votre État là État s 0S un et s 2 c'est. Je suis devant quelle machine quoi ? Quand j'active quelle machine je peux prendre la décision de d'aller activer la machine un, 2 ou 3 ? Mais je vais pas forcément gagner à chaque fois quoi. Donc je prends une décision, je je choisis une action, mais l'action peut mener vers d'autres d'autres environnements. Ou je sais pas. Bah explique moi un jeu de rôle, les jeux de rôle vous allez choisir de faire une action je sais pas je tire sur Intel. Euh. Ou dans le jeu vidéo j'en sais rien. Bah on prend une décision de faire une action, mais on n'est pas certain que que l'action va réussir. Donc ça peut nous emmener, ça peut faire, donc ça peut donner lieu à des actions risquées par exemple. Je prends le risque de de me de sortir. Je sais pas s'il va pleuvoir, j'ai surtout pas envie d'être mouillé mais je dois aller absolument en cours donc je prends l'action d'aller en cours. Mais je sais pas si je vais me retrouver dans l'état sec ou dans l'État mouillé parce que ça va dépendre de ça va dépendre de si il pleut ou pas quoi. OKOK. Hop. Donc. Là on garde les mêmes définitions. Bah ça normalement vous l'avez déjà, vous l'avez déjà noté, mais c'est très important de bien bien être à l'aise avec ça. La probabilité de transition probabilité de se retrouver dans les dans l'État esprit AT +1 sachant qu'on était à s en t et qu'on a pris l'action a ? Ouais. De l'environnement ? Bah typiquement la probabilité des actions ça va être toi qui va choisir. Médine environnement échange qu'on trouve les non, l'environnement il. Enfin quand tu peux tout changer si tu veux, mais dans dans le cadre du renforcement learning, ton environnement est fixe. Et nous justement, ce qu'on va chercher c'est trouver la la meilleure action à faire à chaque fois pour maximiser notre réponse. Ok. Y a une question là bas ? Ensuite ? La fonction de récompense donc c'est pareil. Sauf que cette fois la fonction de récompense sachant que c'est le c'est l'expérience de ma récompense à t'épuise un maintenant elle va aussi dépendre de mon action. C'est bon ? Parce que comme l'état de l'État n'est +1 dépend de mon action, forcément ma vraie. Ma récompense va aussi dépendre de voiture. Donc maintenant c'est plus R de s, C'est R de SAOK. Et donc encore une fois, quand je l'écris comme ça, petit à petit a c'est une fonction déterministe. Et si je l'écris avec des grandes variables, c'est que on est sur quelque chose de stock. On retrouve les définitions de de politique déterministe et politique stochastique. Ok, on peut avoir un agent qui décide toujours de faire la même action. Quand il est dans le même État. Donc je suis dans l'État, je suis dans l'État classe 2, je sais que dans tous les cas j'ai à l'examen ou alors je décide d'avoir une politique qui est stochastique c'est-à-dire hop en fait non ce soir j'ai pas envie je vais aller au pub et j'y retournerai encore de Marie. Ok ça politique stochastique. Et encore une fois je vous ai déjà expliqué, on fait bien la différence si on voit qu y a juste un petit test dedans, on sait que déterministe, si on voit qu'il y a sachant s, Ben on sait qu'on travaille avec des probabilités, et donc on fait attention parce que il y a des espérances qui rentrent, et cetera, et on fait attention à ce qu'on manipule. Bon voilà, du coup ça c'est clair pour tout le monde politique déterministe et explique. Ok. Maintenant, on retrouve notre notre exemple de l'étudiant. Je. Cette fois, on a des décisions. Donc quand on est dans la classe une, on choisit d'aller sur Facebook, on choisit d'aller études. Sauf que ici, ce qu'on remarque c'est que on est dans le cas que je vous expliquais tout à l'heure. Ou. Notre action est déterministe. C'est à dire que si je choisis des Facebook, je vais aller sur Facebook. Personne m'a empêché si je fais. Donc la seule différence avec. La seule différence avec. Avec le graphe précédent, c'est que vous avez remarqué. On a ajouté un peu un un label sur les flèches. Pour dire que une flèche est associée à une action. Ok. Et donc là, on avait juste des probabilités de transition, de d'aller de l'un à l'autre. Et maintenant ? C'est plus une probabilité de transition, c'est une action ça, c'est c'est à dire que c'est le modèle qui choisit par lui-même de faire telle ou telle. Donc typiquement, si on revient dans dans ce dans ce cette ici. On peut imaginer que en fait le 0 5, typiquement quand on a la classe une, la probabilité d'aller d'aller sur Facebook de 0 5, ça peut être la policy de notre modèle. C'est à dire que le modèle il peut choisir et dire OK si je suis dans la classe une bah une priorité 0 5 pour que j'aille sur Facebook, une probabilité de 0 5 que j'avais dans le. Ok, donc tout se recoupe un peu, mais c'est juste que maintenant on parle plus de. On a plus juste un un processus de décision. Pardon on a on a +1 MP où on n'a pas de décision. Là on subit le changement juste à chaque fois on tire les dés et on voit, on voit ce qu'on fait là. Cette fois, c'est à nous de choisir. Enfin, on a la main sur le le, le mécanisme de décision de l'agent. Et on va pouvoir choisir de mettre une probabilité 0 pour Facebook par ex. Pas facile, ouais. On tire les 2. Et bah là justement, tout le but d'une force. Mon planning c'est trouver le meilleur mécanisme de décision possible. J'ai mon robot, j'ai accès à mes, à mes moteurs, je suis dans mon labyrinthe, je peux aller dans la que je veux. Bah le problème Dunning, c'est en fait t'es dans un État, t'as une observation, une récompense, ce que tu fais. Le RH, du coup, c'est une décision humaine. Ouais R le chef, du coup on peut en parler si vous voulez, mais c'est RH c'est ce qui est utilisé pour fine tuner des modèles style chat GT où on va le modèle il a appris à faire du bullshit sur plein texte et on va le fin tuner pour que ce qu'ils disent ressemble plus à ce que ce qu'attendrait l'utilisateur. Parce que RH c'est pour une force planning, c'est ce qu'on va utiliser, ce qu'on fait là from Human feedback donc ça veut dire c'est l'humain dit ce que tu dis là chez GP c'est bien, là c'est pas bien, là c'est bien c'est pas bien. Donc concrètement ça veut dire que quand on dit bien ou pas bien ça lui donne une récompense. Donc le le GP est récompensé. Quand on lui dit c'est bien ce que tu dis, il est puni quand on lui dit C'est pas bien. Et donc il essaye de s'adapter, de trouver la meilleure progression, de trouver le le. Il adapte sa politique pour prendre l'action qui vaut bien et donc l'action de Gpt c'est le texte qui va produire et donc chaque GPT optimise sa politique. Pour avoir la meilleure récompense, c'est-à-dire le meilleur feedback utilisateur. Ce qui est logique finalement, on optimise cette GT pour qu'il réponde de la meilleure manière à ce qu'on dit. Voilà. Je toc, toc toc, on est toujours dans dans les MDP, on garde les mêmes définitions. Sauf que cette fois. Function, il y a un petit PI. En plus, qu'est-ce peut m'expliquer pourquoi ? Vous en parlez rapidement tout à. Pourquoi il y a pas de PI solution à droite là ? J'ai passé le RTD. Ouais. Enfin il y a ouais c'est ça, y a pas de tu prends pas de décision. Tu ne choisis pas la décision ? Ouais non mais c'est exactement ça parce que la value function garde la même définition. C'est l'espérance du gain, sachant l'État, l'état initial. Sauf qu'en fait maintenant, vu que je choisis enfin vu que l'agent va choisir la décision qu'il prend. Bah en fait, si on a, si on. La value function, elle va dépendre de ce qu'il choisit de faire. Si vous dans le labyrinthe typiquement et que le modèle il choisit. On a trouvé une politique nulle ou le modèle, il tourne en rond. En fait ma value function elle sera elle sera moins l'infini partout. Parce que l'espérance de mon gain. Elle est égale à l'infinie sachant qu'on sort jamais du labyrinthe, donc en fait Mary Ward à chaque fois, c'est toujours moins 1-1-1-1-1-1OK. Sauf que si j'ai la politique qui optimale donc la politique me fait toujours sortir du labyrinthe. Alors dans ce cas-là le la value function c'est celle qu'on a vue où on avait. On avait un truc décroissant jusqu'à moins un à la sortie de la. Ok, donc maintenant qu'on a des décisions de la part de l'agent, bah en fait la tension nécessairement elle dépend de cette politique là. C'est clair pour tout le monde, mais ça garde la même définition. Donc c'est faut juste garder en tête que maintenant qu'on a une politique, maintenant que l'agent il choisit de faire des choses dans dans l'environnement. Eh bien. Eh bien la la, la value function dépend de de cette politique. Et donc nouvelle définition maintenant. Qui est l'action ? Manifestation pareil, ça a très bien connaître. Je sais pas si vous c'est la fonction Q je sais pas si le terme de Q learning learning, Deep learning ça vous parle ? Et Ben le Q, il vient de sortir. Ok. Donc la Q la cuve elle enfin la cuve function. Donc là encore une fois plein de termes différents. Mais vous pouvez noter à côté action value function function cut table. Il y en avoir quelques autres. Donc là 3 termes qui veulent dire de la même, qui veulent qui, comment dire qui qui réfère à la même fonction, qui est cette fameuse fonction Q de S et a ? Ou simplement ? Maintenant qu'on a des actions, on peut aussi choisir d'indexer la value function sur l'action. C'est-à-dire ? Quand on est dans l'état ici. Eh bien. C'est pas un excellent exemple, mais quand on est dans l'état classe un, on a 2 options. On peut soit choisir d'aller sur Facebook, soit choisir dans d'aller dans la classe de. Euh. Ce qu'on avait fait précédemment c'est qu'on avait calculé la value function pour chacun de ces États, c'est à dire c'est quoi l'espérance de l'espérance de de gain ? Donc gain c'est quoi ? C'est la ? C'est quoi le gain ? Ouais, la somme que la somme pondérée de toutes les récompenses. Très important. Et vraiment ces définitions là. Faut que ce soit très clair pour vous tout de suite. Parce que les termes que j'utilise enfin ouais, faut savoir de quoi on parle. Donc quand je suis en classe une j'ai 2 options. Et précédemment on a calculé la la value function, la value function elle prend pas en compte les actions. Elle me dit juste que si t'es dans cet État là c'est bien c'est si t'es dans cet État là c'est pas bien. Maintenant on a des décisions à prendre, donc en fait ce qui serait bien c'est qu'on a une fonction qui me dise si c'est bien de quand je suis dans un État s, si c'est bien de faire l'action a ? Ce serait mieux, non ? Pour choisir quelle action faire à chaque fois, OK. Donc maintenant qu'on a des maintenant, qu'on a des décisions à prendre. Eh Ben on va rajouter un paramètre à la fonction V qui nous donne la fonction Q. Qui est la fonction Q et après en entrée, l'État où on est et l'action qu'on fait ? Et elle me renvoie l'espérance de mon gain sachant ça. Ok, donc c'est exactement la même définition, enfin la même manière de définir leur fonction, parce que cette fois on indexe par l'action e par l'État et l'acte. Ok, ça veut dire que d'un point de vue informatique derrière. Une fois que vous avez la fonction Q. Et Ben en fait, à chaque État vous allez pouvoir calculer le cul de chaque action. Vous pouvez faire et trouver par exemple la meilleure action à faire sachant que vous êtes dans l'État a. Ok donc dans cet État là typiquement vous allez voir que OK, je suis dans l'état class une. Je connais plus de classe une aller sur Facebook, l'état de Q aller à la classe 2, donc je choisis d'aller sur Facebook. Ah non du coup j'ai mal appris ma fonction mais mais ça me permet de prendre la décision en. Ok, et donc toutes les méthodes de to learning, et cetera. En fait c'est ça, c'est apprendre ma fonction Q, c'est à dire ? Apprendre à discerner si dans un État donné, quelle est la valeur de l'action que j'apprends aux échecs. Je suis dans un État particulier du de l'échiquier. Quelle est la valeur de l'action de bouger le cavalier dans telle de telle position ? Et cetera. OK, très très important comme notion la fonction Q, donc action value. Fonction, voilà. Hop je oui un peu la, franchement ouais euh là j'ai vu ça aussi en paramètre la politique. Comment j'ai vu ça prend aussi en paramètre la politique c'est ouais, remarque judicieuse, c'est comme la value function, encore l'action vu que avec la politique, l'action. Euh. C'est une bonne question. Euh, ça va dépendre oui, si. Si, si, parce que donc, excellente question, la question c'était pourquoi il y a encore un PI en paramètre du cul, sachant que cette fois on prend en compte l'action ? Et donc la réponse c'est la même en fait que le pour la version précédente. Parce qu'en fait là tu prends l'action à à a et tu prends pas en compte toutes les actions futures. Donc pour connaître la l'espérance de ta reward à la fin il faut bien connaître la il faut bien connaître toutes les actions que tu vas prendre mais mes remarques. Euh TIC TAC, donc là ça vous donne un exemple de de l'action value ? Euh, pour une politique donnée. Donc ici, on est dans le même exemple et. Et on a quasiment. Mais on a pas exactement les mêmes, les mêmes étapes. Mais donc là ce que ce que je vous donne c'est si jamais. Ma politique ? C'est 0 5 partout, c'est à dire je fais, je fais n'importe quoi, alors ça c'est ma ça, c'est ma vie future. Ok. Donc là elle a moins de sens en soi vu qu'on prend pas de décision et donc ça donne une idée de de ce que ça peut, de ce que ça peut donner. Maintenant. Euh, on va passer rapidement dessus. Je pense qu'on a fait assez de maths tout à l'heure. Euh, on généralise ce qu'on a vu avec belman précédemment. Franchement, c'est exactement pareil, sauf que au lieu de de faire juste en fonction de s, on rajoute un conditionnement par a. Ok donc le calcul est exactement le même, sauf que vous allez donc vous allez utiliser fubini encore là-bas là fubini a bien connaître. Lire les. Ce truc là ? Hop. Hop. On aime les maths en. Mais c'est important. Donc on fait exactement la même chose. On obtient la l'équation de Bellman pour la fonction Q quand on index par l'État et l'action. Donc. Quand on développe. Donc quand on développe l'espérance. C'est exactement ce qu'on a fait tout à l'heure. Eh Ben on obtient la forme explicite. De la fonction de valeur. Vous aurez les slides, hein, pour ceux qui qui recopient. Donc quand on développe, on fait l'espérance de tout ça. Donc c'est quoi l'espérance ? C'est si je suis dans l'État s bah en fait ça va être la probabilité de prendre l'action a sachant qu'on est dans l'État est donc la \*\*\*\* de a sachant SOK, ça c'est la probabilité multiplié par ensuite. Donc là c'est comment dire ? On utilise l'équation de Bellman, donc vous reconnaissez le la récompense à l'instant t plus gamma fois la vallée fonction à l'État suivant. Et ça on fait juste pour la somme des transitions si je sais que je fais l'action a ? Et je suis dans l'état s c'est quoi les les États possibles dans lequel je peux me retrouver ? Ok donc ça c'est franchement l'application bête de. De cette équation là. Dans le cas de d'un MDP. Ok, ça vous pourrez je pense, c'est un bon exercice de vérifier chez vous que vous arrivez bien à ce résultat là. Vous partez de l'équation de Bellman. Et vous vérifiez que vous arrivez à écrire ça ? Ok. Et on le fait exactement de la même manière de ce que son actif ici. Ok, on calcule l'espérance. L'espérance c'est quoi ? C'est je sais pas sur tous les cas possibles. Tous les cas possibles, c'est quoi ? C'est mes actions et ensuite je sépare sur toutes les les cas possibles une fois que j'ai pris une décision sur l'action. Et ça c'est la probabilité de transition de d'aller à exprime sachant que j'étais en S et que j'ai pris l'action là. Ok. Ok et en dessous ensuite c'est écrit pour la même chose pour la. Pour que. Voilà. Le. Moi je sais pas quoi trop dire de plus. Euh. C'est pareil, c'est la définition et vous utilisez bellmann aussi pour. Pour développer, pour développer la forme qu'on a là, et donc ces ces équations, elles sont importantes parce que c'est un peu la base de. Ce qu'on va programmer quoi ? Vous pouvez pas programmer une espérance ça se fait pas. Par contre les programmer une somme ça vous pouvez faire. Du coup quand on écrit comme ça c'est plus facile après en python de faire sa boucle pour dire OK je dois calculer la somme de tout ça donc je fais OK fort action possible action OK je fais la somme fort tous les états dans tous les états possibles On va pouvoir rétro propager. Sur sur l'établissement. Ok. Très bien. Bien sûr. Ouais. On va finir ça et en pleine pause, ouais. Il reste ouais 3 slides. C'est bon, pas de question là-dessus. Franchement, bon exercice de vérifier que vous arrivez à écrire ça, c'est important. Ok. Donc là, c'est un peu ce que je viens de vous expliquer. Sur comment est-ce qu'on fait pour écrire le fameux produit ? Où on a l'état s ? On a les les actions possibles en fonction de en fonction de. De mon état. Donc ça dépend de mon environnement. Quand je suis dans l'État je peux choisir de je peux choisir de prendre l'action a ou l'action a prime. Et ma value function. Ça va être donc l'espérance de de mon gain. Et que j'écris avec bellman comme là le l'ensemble donc la la probabilité de pardon, l'ensemble sur toutes les actions possibles de ma fonction Q que je connais déjà sur l'État d'après. Avec la probabilité de passer à l'autre avec que je connais grâce à ma politique. Ok, c'est juste la même chose que que ce qui est ici. En reconnaissant la forme de Q. Le s qui est dans la parenthèse elle est grande ou ou c'est impossible ? C'est un petit et du coup c'est un petit s ouais parce que justement on a plus d'espérance maintenant. Tu vois, on a plus que des petits parce que on est passé. On est passé de ça. Là, on a l'espérance d'un grand air, d'un grand et d'un grand a. Et maintenant qu'on connaît exactement les actions possibles des États possibles, on passe en petit test s. Et petit a. Et on en fait la somme pour calculer l'espérance avec l'idée que mon grand a qui m'a variable aléatoire va prendre la valeur petit ou petit à prime en fonction de ce que je choisis de faire et de m'appeler. Et donc ouais là c'est des c'est des formules que vous pouvez implémenter en Python. Toc Toc, ça, c'est juste la même chose, écrit différemment. Euh. Ouais. Pour moi ça on l'a déjà fait, je sais pas si elle vous parle de l'avoir paragraphe ou pas. Mais ça vous donne, c'est un peu ce qu'on ce qu'on a fait tout à l'heure avec. Avec le petit calcul de l'application. Ou on connaît. On connaît la value function pour tous les États possibles après s. On sait. Qu'on a 2 actions possibles. Et donc pour calculer avec bellman pour calculer la value function à l'état S, et Bah je fais mes 2 sommes. D'abord je somme sur les actions possibles et ensuite je somme sur la probabilité de passer de de s prime sachant S et a. Multiplié par la valeur qu'on connaît déjà de l'état d'appel. Est-ce que ça c'est clair ou pas ? Je pense que le GRAPHE est pas mal en vrai pour se rendre compte de ce qu'on est en train de faire pour calculer l'espérance. Mais ouais, t'as une question. Les États expriment là ouais et en fait on peut rétro. Exactement. Ce qui se passait avant, après, ça nous sert en fait pour. Bah c'est la question de l'exploration. C'est ce qu'on disait tout à l'heure dans le cadre du labyrinthe. La, la fonction de valeur, on la connaît à posteriori parce qu'on a tout exploré. Dernière étape ? Exactement. Et ça permet de la prochaine fois que tu en labyrinthe, en fait tu prendras directement la bonne direction. Il y a un peu l'idée de la ce qu'on fait, on est sur des trucs basiques où on on apprend rien d'intelligent, juste on traverse tous les États possibles et juste on les retient intelligemment pour pouvoir se dire Ah bah la prochaine fois que je suis dans cet état là, la bonne décision à prendre c'était d'aller dans telle direction. Ok, et et c'est pas évident de le faire. Parce que si vous avez fait 1000 étapes dans votre labyrinthe, c'est pas évident de vous dire, Ah mais je suis là, je suis arrivé. Donc ça veut dire que à l'étape 500, enfin l'étape 502, il fallait que je tourne à droite dans le cadre du même labyrinthe. Le même environnement ? Exactement exactement. Ouais, on fait des trucs très basiques pour l'instant. Ouais, je suis d'accord qu'à la fin on aimerait avoir des robots un peu plus malins que enfin justement quand quand on change un peu l'environnement, que le truc se comporte bien, mais là on est vraiment sur des des labyrinthes basiques où le modèle dit OK, là je suis en 2.2, il faut que je bouge en : un. Ok. Donc ouais je pense c'est pas ouais c'est c'est j'aime bien écrit comme ça ça permet de enfin c'est vraiment exactement la même chose que ici hein. La somme sur a la somme sur s la probabilité de transition, la reward plus gamma fois la fonction. C'est exactement pareil, simplement, on l'écrit différemment pour. Avec ce petit graphe en tête donc ça c'est pour vous aider à retrouver cette formule là OK ? Ok, et là c'est pareil pour la fonction Q. Sauf que cette fois, du coup, nous, ça nous intéresse, c'est. C'est de savoir pour l'action, pour une action en particulier. Là on veut regarder. C'est quoi le meilleur État suivant là on veut regarder. Je sais que je prends déjà cette cette action. Et donc je regarde un cran plus loin, c'est-à-dire que je regarde l'état suivant et l'état suivant. Avec une action, c'est-à-dire que la, la fonction Q, c'est pas juste l'État, on a vu, c'était l'étape plus l'action, donc c'est un peu je regarde si ça c'est bien, si ça c'est bien, si ça c'est bien, si ça c'est bien. Et de la même façon qu'avec les questions Goldman, je rétro propage pour dire je sais déjà Q pour Q et a. Je sais déjà quelle valeur j'ai donc j'arrive à calculer pour mon État et c'est action quelle est la valeur de ma queue ? Ok pareil, du coup vous pouvez vérifier que vous arrivez à. Que vous arrivez à écrire ça ? Va y reste un concept dont je vais vous parlerEt c'est là la cause du. Donc du coup on se dit 20 Min 1/4 d'heure ? 15 Min. Allez, rendez-vous à 10h05 à 10h10. Où la machine à café 3e, Ah oui, parfois il marche pas. Ils sont même endroit sur les étages. Non là là là. On est présent. À la 2nde près, il faut que tu rentres dans la fille. Et après moi j'ai pas le droit à ce cas-là. J'ai envie de savoir. On s'assure que tout est bien consolidé parce que là, ce qu'on vient de faire, je sais que ça fait beaucoup de choses, mais c'est vraiment hyper important que vous maîtrisiez toutes les toutes les notions qu'on a, qu'on a posées. L'équation de Bellman. Enfin le le quand quand j'écris 1S 1AO c'est hyper important. Je vous parle de gain, d'évaluer function, de quelle fonction il faut que faut que vous sachiez à chaque fois de quoi je vous parle. Sinon je vais vous perdre tout de suite dès qu'on va commencer à faire les choses plus compliquées. Donc là on va consolider ça ensemble. Je vais vous montrer rapidement sur les slides 2 3 exos pour vérifier que vous avez bien compris. On parle d'un dernier sujet et ensuite je pense qu'il répond un peu de temps. On j'aimerais bien. Bah je les ferai. Ceux qui ? Ceux qui doivent partir tout de suite, mais qu'on discute rapidement de comment vous avez senti le cours et si y a des trucs que vous aimeriez faire en particulier, peut-être ça, ça vous a remémoré des choses dont on a parlé. Des sujets plus intéressent, j'entendais des gens qui parlaient de voyagers là. Je peux éventuellement faire des séances où on parle de ce genre de modèle pour voyager, c'est le c'est un papier de recherche est sorti l'année dernière qui où des gens vont entraîner. Il n'y a à à finir Minecraft. Et donc c'est un modèle de RL derrière donc sauf que c'est c'est bien que ça peu plus compliqué qu'un modèle de Markov où qui apprend une table quoi. Donc voir un peu voilà qu'est ce qui vous intéresse s'il y a des choses que des choses dont vous voudriez que je parle parce que bah derrière moi je peux aussi adapter le les derniers cours pour que. Bah pour faire des trucs vous intéressent. Voilà voilà. Du coup je vous laisse, je vous le dis maintenant pour que vous pensiez déjà à quelques questions, suggestions pour la fin, mais on en parlera à la fin du coup. Logiquement voilà OK alors du coup, rapidement. Des petits exercices là donc les slides seront sur moodle vous pourrez vous pourrez regarder on ça on l'a un peu déjà fait donc je pense que ça a rien qu'on le refasse ensemble. Là c'est le genre de trucs que je pense que je mettrai à l'examen final. Euh, sur la partie maths, après y aura plutôt une partie, une partie sur ce qu'on va faire plus tard, mais c'est vraiment à bien maîtriser. Et pour demain, si vous voulez comprendre ce qu'on fait en TP, Ben il faut vraiment maîtriser des trucs. Donc voilà les petits exercices. Des petits exercices et. Ouais. Un concept ? Un concept dont on va enfin qu'on va utiliser demain, c'est un principe d'optimalité. Et donc là on se replace dans le cas où on a. Qui, cette fois, dépend de la politique ? Ok. Oui. La question c'est bah comment est-ce qu'on trouve la meilleure politique quoi ? Est-ce que vous avez des idées sur comment est ce qu'on peut faire si jamais si jamais j'ai ma politique ou ma value function genre à partir de l'un ou l'autre comment est-ce que je fais pour trouver si j'ai ma value function disons comment ce que je trouve un politique ? Sur les 2, les 2 sont liés comme on a vu puisqu'on a un index. Mais comment est-ce qu'on peut arriver à lier la meilleure des politiques d'une certaine manière avec la value ? Je sais pas, tu peux parler plus fort. Ouais, enfin ouais. On va cherche à trouver un maximum, c'est bien d'autres idées. Non, OK. Quand on parle de cette formule là avec l'espérance. On a quelque chose de donc qui est déterministe. Sauf que si on se transforme dans un dans un environnement type labyrinthe et qu'on a une politique qui est fixée, et on suppose que cette politique là c'est la meilleure des politiques possibles. Donc on sait que dans tous les cas on fait le meilleur Move qu'on peut faire dans la vie. Alors en fait l'espérance elle saute. Puisque tout est d'accord, il y a plus d'espérance, on a plus de caste cité à chaque fois notre politique est déterministe, donc on a plus de s directement, on a plus, on a plus de sachant États, et cetera. Et notre histoire paraît, elle est fixée, et on a, on est dans un cas, le labyrinthe où le le modèle d'action il n'est pas stochastique se déterministe. Si je décide d'aller à droite, je vais à droite et personne va m'en empêcher. OK ? Donc en fait dans ce cas-là, donc dans le cas plutôt simple du labyrinthe, on a plus de stock de stock. Intimité, pardon à ce moment-là. Euh. Et donc Ah non bah ici pardon, ici on regarde la stochasticité mais dans ce cas-là on va voir que donc ce sera pour demain, on aura même plus de probabilité intérieure. N'empêche que on va noter, en supposant qu'on l'a, la value function optimale qui est obtenue pour la politique la la meilleure qu'on puisse obtenir. Donc ça, ça veut dire quoi, ça veut dire que euh. Pour chaque action qu'on a, enfin pour chaque. On se remet dans le dans le cadre de branchement ici là. À chaque fois qu'on va avoir un choix entre les actions à faire, on sait qu'on fera toujours la même, la meilleure action. On est d'accord puisque puisqu'on a la meilleure des politiques. Donc en fait on va, on va plus avoir la somme des probabilités de la politique action selon s ici. On va juste le Max très bien. Donc. C'est là, non ? Donc on n'a plus la policie de de l'action en fonction de l'État on a plus on a plus qu'un Max. Donc là on garde toujours une somme sur les étapes possibles de s prime avec la probabilité de transition. Mais dans le cas du labyrinthe du coup cet exercice on fait sauter la probabilité et on garde uniquement l'état qu'on obtient si on fait l'action. C'est bon et donc là typiquement si on fait sauter la on a une fois la capacité. D'ailleurs ça vous pouvez le programmer en python. D'accord, vous pouvez calculer la somme, vous connaissez les probabilités de transition simplement dans le cas c'est pas Stochastique y aura juste des uns et des zéros mais ça ça marche quand même OK ? Donc en fait. Là, ce qu'on vient de ce qu'on vient d'avoir, c'est d'avoir une une condition pour avoir la value function optimale en fonction de la value fonction optimale sur un autre État. Donc on a un moyen simple de vérifier. En supposant qu'on a la version la value function optimale, on vérifie que. Euh, on peut vérifier que les biens optimale de manière. Donc on sait que si on a la politique optimale, alors la fonction ici s'écrira de cette manière-là et on pourra le on pourra le vérifier. Et donc en pratique ça veut dire que si vous avez la la value function optimale partout sauf à un point. Alors vous pouvez le déterminer grâce à grâce à cette équation là exactement comme tout à l'heure. OK là ce qui change du coup c'est que. On on se place vraiment dans le cadre du renforcement learning et. On fait vraiment l'hypothèse de en fait la politique je l'ai pas donc la value function je l'ai pas du tout vu que la value function dépend de la politique. Et donc on se dit OK. En supposant que j'ai la politique optimale, alors j'ai une value function optimale associée. Et là grâce à belman, j'ai une équation qui la lie et qui va me permettre d'obtenir des informations dessus. Voilà et là on ce qu'on sent un peu aussi à ce que je viens de vous dire c'est que Ben en fait, même si j'ai une value flunch optimale partielle, bah si il manque plus qu'un point, je vais réussir à à le le détecter enfin à à à le compléter OK ? Et donc en pratique c'est comme ça qu'on va programmer des méthodes pour pour trouver cette fameuse value function imal, ce qui est équivalent à trouver la policy optimale aussi on est d'accord, c'est clair pour tout le monde hein. Si je trouve la value fonction optimale, j'ai trouvé la politique optimale. Ouais et inversement tout à. Euh OK donc donc les notions à bien à bien connaître c'est juste on a une value function optimale qu'on note westar, et pareil pour la queue function custard OK, et on note aussi, vous pourrez aussi trouver un 1Q, un petar, si pour pour la politique optimale. OK, c'est comme ça qu'on le note et puis. Et donc ce qui à retenir, c'est que grâce à l'équation de Bellman, on a des. On a des moyens, des moyens de de de lier ces notions entre elles pour trouver l'une à partir de l'autre ou pour compléter une connaissance partielle qu'on aurait sur l'une ou l'autre des fonctions. Ok, ici pareil à faire chez vous. On vérifie, on on suppose. Donc là on on garde, on reste dans le dans le cas où on a un gamma de un et une des probabilités de transition qui sont égales à une en fonction de de notre action. C'est clair pour tout le monde. Si on fait la, si on fait le choix d'aller sur Facebook, on y va et on n'est pas. On n'est pas emmerdé. Je. Là je vous donne la value function et donc on vérifie exactement comme tout à l'heure. Euh. Oui bon là là je vous mets la la comparaison entre, fonction et que fonction. Donc vous voyez que c'est pas exactement la même chose. Dans un cas, on va pouvoir choisir, OK, je suis dans l'état. Je suis dans l'état classe une. Et je sais que du coup le meilleur État suivant c'est bien d'aller en classe 2 et pas d'aller sur Facebook, on remarque bien que la la la. La CUE enfin la la valeur de de la queue function à ce moment-là c'est pas exactement pareil. C'est-à-dire que on va avoir une une une fonction Q associée à je suis en classe un et je vais en classe 2 qui est 6. Et je suis en en classe une et je vais sur Facebook de 5 et c'est pas exactement pareil que la value function en en la case Facebook. Donc c'est des choses qui sont très liées comme on l'a vu, mais c'est pas la même chose non plus en termes de de valeur numérique. Donc ça vous donne un exemple. Et ça, c'est TAC. Ok Ouais, je pensais qu'il y avait. Non ? Ok, ce sera après. Euh très bien. Et donc associé à ce custard et à ce à ce westar. On a la fameuse politique optimale qui. Comme on l'a vu en fait, si on a la fonction Q ou la fonction V. On l'a directement, c'est-à-dire qu'on prend juste la meilleure des actions possibles sachant l'étalon est. Et là, on a notre politique optimale. C'est clair pour tout le monde, c'est assez. C'est assez simple comme idée. Si on sait dans tout moment c'est quoi l'espérance de de mon gain, si je suis dans l'État et que je le l'État a, il suffit de choisir l'espérance, le la plus grande, pour maximiser mon gain. D'accord ? Tac TAC TAC, donc. Ouais, et dans ce cas-là, du coup la politique elle est même plus stochastique devient déterministe, même si on l'écrit ici de manière stochastique, même probabilité de un donc en fait on sait que on va être tout le l'action optimale. Et une propriété intéressante, c'est que pour un processus de décision, markovien le MDP ? Donc je rappelle, c'est bien l'ensemble. État probabilité reward, action, on a les 4. Donc pour 1MDP. On a toujours une une politique optimale. Ok, donc si vous avez un modèle style ? Ce truc là, hein ? Vous savez qu'il existe une policier optimale. OK vous avez moins de la trouver algorithmiquement pour des pour des si votre graphe est pas trop gros, OK ? Dans la vraie vie, c'est pas vrai, mais vous savez que si vous avez un cas très simple comme ça, alors. Vous avez l'existence de la de la police optimal. Ok alors donc là là je vous ai mis des petits exercices à faire chez vous. C'est voilà. Si vous les lisez, vous vérifiez que vous vous êtes à l'aise avec les notions qui sont dedans. C'est le genre de question que je qu'on peut mettre à l'examen. Voilà. Voilà je vais je pense, ça sert à rien qu'on détaille plus que ça. Toc, Toc, Toc. Ok, donc la dernière petite notion. L'annexion, la notion d'exploration et d'exploitation. Et est-ce que ça parle à certains ou pas ? Ouais. Vous l'avez vu ? Avec quoi ? Ouais, très bien. L'épisode crédit, ça vous parle aussi. Vous avez tous fait ce cours là où c'est le cours de l'année dernière, c'est ça ? Ok bon bah du coup ça va ça va vous parler. Donc c'est vraiment en conceptuel. Souvent, en termes de de programmation, vous allez choisir la manière dont on va. Votre votre agent va se comporter pour explorer l'environnement comme comme on a vu dans le cas dans les cas simples, on se place là, on va tout explorer. Donc on fait que de l'exploration. Et ensuite il y a une partie exploitation c'est à dire une fois que je sais enfin une fois que je connais ma value function et Ben en fait je peux juste exploiter la connaissance que j'ai pour faire tout le temps la trajectoire optimale. On est d'accord ? Et donc en pratique c'est jamais l'un ou l'autre, c'est à dire que dans des dans des espaces trop grands, si vous faites que de l'exploration vous avez jamais trouvé votre votre chemin. Donc à un moment donné il faut vous dire OK, dans cette zone là c'était pas trop mal. Donc je vais rester pas trop loin et je trouverais peut-être de meilleurs des meilleurs endroits. C'est exactement comme la vidéo que je vous ai montrée au début là avec le le robot qui apprend à marcher. À la fois de l'exploration, c'est à dire que le robot va essayer de des nouveaux, des nouveaux gestes, mais à la fois une fois qu'il a trouvé un truc qui marche, bah il s'y tient quoi donc au début, là, quand il faisait le quand il faisait la limace. Là on parlait d'exploitation, il a trouvé un truc qui marche, c'était tellement dur de trouver un truc qui marche. Il est resté là-dessus, il est-il est-il a pas tenté de de faire d'autres choses donc il a fallu l'aider en lui disant OK maintenant tu vas, tu vas devoir tenir debout et tu seras puni quand tu vas taper le sol pour que il se décide finalement se mettre debout. Et cetera. OK donc. Voilà, c'est c'est des, c'est des termes flous. Mais souvent on on va en parler et et c'est un, c'est bien d'avoir ça en tête quoi. Donc voilà par exemple, quand on est dans un un dans un labyrinthe. Eh Ben on peut choisir. On peut choisir d'explorer ou de de d'aller dans la direction où on sait que c'est à peu près une bonne direction, sachant qu'on on. On peut se tromper aussi quoi ? Euh. Un autre exemple, c'est les systèmes de recommandation. Euh, pour enfin je sais pas, tous les réseaux que vous utilisez sur Youtube, et cetera, le système va à la fois vous proposer des choses qu'il sait que vous aimez. Euh sauf que parfois vous avez remarqué que y a un poste ou une vidéo qui est un peu différente quoi. Parce qu'en fait le modèle se dit bon en fait on va tester ça, peut-être que ça va lui plaire. Parce qu'en fait si je sais pas. La première vidéo vous Regardez sur Youtube, une vidéo de chat sur Youtube vous repose. Que des vidéos de chat, vous Découvrez jamais les vidéos de chiens. Alors qu'en fait, peut-être que ça vous plairait quoi ? Donc il y a une y a cette notion là aussi dans les systèmes de recommandation si ça vous intéresse. Qui est. Bah à la fois il faut exploiter ce qu'on sait sur les utilisateurs. Il aime les chats à la fois. Il faut savoir explorer de nouvelles possibilités pour découvrir de nouvelles choses que sinon en fait on fait jamais de découverte. Me voilà sachant que c'était la petite image rigolote que je mettais ici. Mais l'exploration, ça peut coûter très cher. Quoi, un robot ? Vous déployez un robot à 3000000 d'euros dans votre labo. Vous avez pas envie que votre robot il explore importe quoi ? Parce que il va se casser en 2, il va tuer quelqu'un. Donc là on va plutôt être sûr de l'exploitation d'un truc, d'une, d'une. Qu'il a déjà appris en offline, c'est-à-dire en simulateur. Par contre en simulateur bah on peut le laisser explorer ce qu'on veut, c'est pas grave. Donc voilà une notion à bien à bien à vendre. Ce donc pour l'exploitation. Un agent de trading. Un agent de trading il va pas faire de l'exploration, on a pas envie de d'investir sur. Sur trifouilly-Les-oies Market on peut dire Ah on va peut être essayer ? Non non un agent trading on l'entraîne en offline et une fois qu'on voit qui marche super bien offline on fait que de l'exploitation et dit OK tu fais le meilleur que tu peux sur le marché et puis on voit si on arrive à faire de la. Euh en une ia qui a appris à jouer à un à un jeu une fois vous allez jouer aux échecs, vous avez pas tenter des nouveaux. En fait vous savez que le coût optimal bah vous le jouez. Donc le laser, sauf si vous êtes une ia qui est censée être un peu moins bonne qu'une que la meilleure, donc vous allez vous allez faire exprès de faire des des cours qui sont moins. Donc donc là on, c'est l'idée que on exploite ce qu'on a déjà. Avec toujours le tradeoff, si vous explorez pas, vous saurez jamais c'est quoi les les bonnes décisions à prendre ? Et si vous exploitez pas, bah vous aurez jamais de bons résultats. Donc voilà, OK, je. Je vous remets mon mail ici, là c'est bon pour les slides d'aujourd'hui. Je vais mettre ça comme ça, non, pas comme ça. Hop hop hop et je prends. De. Hop. Ok, donc du coup est-ce que vous avez déjà des questions tout de suite ? Sur la ce qu'on vient de faire aujourd'hui. Non ? Ça veut dire j'étais super clair. Ok, Euh. Je vous partage cette fiche là, c'est pas moi qui l'ai faite, c'est pour ça qui est bien. Ce qui résume un peu tout ce qu'on vient de voir. Il y a toutes les définitions, il y a. Donc la définition d'un MDP typiquement là parce que c'était pédagogique, on est passé par le MP, le MDP en vrai on utilise que le MDP on perd besoin de pas besoin de connaître exactement comment ça marche. Mais l'idée c'était de voir. Bah petit à petit j'ajoute une nouvelle complexité et on voit, on voit ce qui se passe. Donc en pratique en R on utilise que des marques dessins. Process, donc y a pas besoin de. En pratique la définition sur les mêmes mais ce qui est important c'est de bien maîtriser le DP oui. Les chaînes de Markov, c'est un tout, un de mes mathématiques qui est beaucoup plus large que le voilà et un peu disons que j'aime de Marco. C'est existe depuis très longtemps et c'est un peu au moment où les gens ont voulu entraîner des des, des modèles à jouer à des jeux, évoluer dans des environnements. Ils se sont dit Bon OK. La définition du du modèle de enfin comment dire du modèle agent, et cetera, elle est assez simple. Comment est-ce qu'on fait ? Enfin c'est quoi le domaine des mathématiques qui nous permet de modéliser ces trucs là ? Et là les chaînes de Markov c'était assez. Donc typiquement le MDP je pense ça existe depuis plus longtemps, mais c'était un peu développé sur mesure pour faire du hein d'une certaine manière, ce qu'on appelle du renforcement de leur niveau aujourd'hui. Mais. Mais voilà, du coup, le processus d'édition Marco vient la définition de ce que c'est qu'une politique OK avec les. Pas forcément très bien mais vous verrez ce sera sur le modèle du coup la définition de la du function avec le gain. Vous voyez, ici y a marqué gain, ça peut aussi être cumulative reward. Enfin, encore une fois, Notez bien tous, tous les, toutes les, tous les alias, quoi qu'on peut utiliser. De. Tac TAC. L'équation de bellman à très bien connaître par cœur, ça va hein ? C'est pas compliqué hein ? Une fois que vous savez ce que c'est qu'un gain, vous savez qu'on peut retirer le premier terme et en fait vous vous revoyez rapidement le la vallée fonction d'à côté. Donc on retient que la value function de s c'est l'espérance de Mary Ward maintenant plus gamma, faudra la value unchain d'après. Voilà donc ça vraiment à bien connaître. Et à bien maîtriser. Pas forcément. Pas très grave si vous savez pas exactement refaire le petit calcul qu'on a fait là avec histoire de fumier, et cetera, mais franchement je pense c'est un plus et ça vous servira pour d'autres cours après enfin je sais pas exactement quel autre cours vous avez, mais si vous avez des cours de probas de de réseau Bayésien. Il y a que ça quoi. Donc si vous savez pas, si vous savez pas faire ça, vous êtes un peu dans la \*\*\*\*\* pour pour faire des trucs plus compliqués niveau programme. Et ce sera beaucoup plus facile pour vous de comprendre quand plus tard il y aura des formules compliquées au tableau, parce qu'on va faire un un modèle de RL compliqué et que j'aurais pas le temps de tout détailler. Bah c'est bien de savoir. Ah OK, là là c'est bien une probabilité, là c'est bien un truc déterministe, là on est sur une politique, et cetera, où vraiment important. Et Ben. Et une preuve de une preuve de bellman, tiens. Donc voilà. Et Ah oui, et ça. On l'a pas vu, mais on le verra demain. Donc c'est pas grave, c'est pas grave OK donc ça je vous le mets sur le moodle. C'est ça, je le mets sur le moodle et les slides. Pour les 2 filles de demain c'est pas possible de l'avoir en avance ou vous voulez le TP en avance justement ? Non je sais pas si c'est possible je peux. Après on n'a pas fait le cours d'eau mais. Il y a quand même une partie, parce que là on a posé les bases mathématiques. Je vous ai déjà donné une intuition sur comment est-ce qu'on allait pouvoir algorithmiquement, trouver les justement, trouver des optimums et cetera, mais on a pas, on n'a pas la théorie. Après là, je vais voir, mais je peux tout tout téléverser sur moodle, je mets tout et si vous intéresse. Faire d'ici demain ? Franchement le TP le TP est pas très long, on aura le temps dans les 4 h de faire largement les 2. Ce n'est pas il n'est pas très compliqué. Ok, est-ce que vous avez des questions là-dessus ? Euh ouais. Ouais. Ce qu'on arrive à représenter un peu sur le monde auquel on est branché, les. Sur des simulateurs, ouais. Ouais et Ben. C'est une très bonne question. Déjà, s'il faut retenir, c'est qu'on arrivera jamais à le faire en. Aussi aussi bien que que la réalité. Donc en pratique tu as des modèles physiques. Donc là je sais pas si vous vous rappelez la petite vidéo là que je vous ai montrée, mais là dans ce cas on est dans un dans un moteur de jeu type unity ou un drive. Il y a aussi Ross, je sais pas ce que vous connaissez. Enfin bref c'est c'est les logiciels, mais dans l'idée c'est un moteur physique où vous allez dire. Okay j'ai j'ai un bras, j'ai un bras et fait telle longueur, il fait telle dimension, j'ai une force que j'applique dessus et le et le modèle il va il va dire Ah bah du coup ton bras il va bouger de 2 cm en ce temps-là enfin et il fait le il fait le il fait le calcul tout seul donc ça c'est ce qu'on va appeler l'environnement et donc l'environnement c'est. C'est ouais, c'est l'environnement quoi. Tu tu lui dis OK mon robot il donc dans le cas d'une voiture, c'était un peu plus simple à comprendre. Dans le cas d'une voiture, votre action ça va être je vais appuyer sur l'accélérateur, je vais tourner le volant, je vais freiner. Et là votre modèle il va dire OK t'appuies sur le fin, c'est pour ça que tu vas arrêter tout de suite. Donc ta vitesse va diminuer de temps. Enfin tu vas être à telle vitesse. Exactement. Ouais, je sais pas si vous avez déjà fait un peu des cours de modélisation 3D ou quoi ? Non ? Bah c'est. Bah typiquement là le la vidéo que j'ai montrée c'est je pense c'est unity, c'est un. C'est un moteur de jeu quoi. C'est vraiment les les les jeux vidéo c'est ça ? Parce que quand t'es programmateur de jeux vidéo tu tu dis OK j'ai un agent, enfin j'ai un bonhomme, il se déplace dans mon environnement et tu lui appliques des forces. Enfin je sais pas si tu fais tirer dessus, bah tu vas appliquer une petite force et c'est et c'est pas toi qui va coder. OK mais si une tire dessus donc. Bon c'est un moteur de physique quoi, c'est de la mécanique là en général t'as un moteur qui fait ça pour toi et donc en pratique pour entraîner des robots comme ça c'est pareil. On recode pas toute la physique, on utilise un moteur qui est déjà présent, qui est qui, qui lui code la physique et nous on va chercher les informations. Donc là typiquement dans la vidéo il a pas codé la physique de de son bonhomme par contre. Il avait accès par exemple à savoir si le bonhomme était debout, pas donc à partir de cette information là, donc s'il est debout, s'il y a des collisions, s'il y a des choses comme ça. Et à partir de cette information là il dit au modèle, Bah OK, ma récompense ce sera moins 3 si il touche le bord, +1 s'il est debout, le la distance a sa target, et cetera. Ce qui est assez Amine, quand ? Parce que de ce que j'ai compris c'est vraiment une révolution dans le jeu vidéo. C'est quoi ce qu'il a apporté par rapport aux autres moteurs qui existent ? Alors question complètement différente mais. Je te réponds rapidement, si tu veux on en reparlera. Mais en gros il là où ils sont meilleurs c'est sur le rendu 3D. Ils ont leur truc des nanas et je sais pas si on a qu'ils connaissent mais ils peuvent faire des rendus 3D hyper hyper réalistes pour pas très cher. Et du coup tu peux avoir un jeu vidéo hyper réaliste en en 60 FPS alors que t'as pas une carte graphique de fou. Voilà voilà. Je t'en prie, d'autres questions. Sur le court, sur les notions, tout est clair. Ok je très bien, je vérifie que j'ai rien d'autre à vous dire. Xbox poste non ? Bah pour moi c'est bon bah Merci beaucoup pour votre attention ce qu'on fait là c'est bon il est 45 je suis très preneur de vos feedbacks donc ceux qui veulent rester, ceux qui veulent partir maintenant y a pas de souci et je vous mets les slides, et cetera. On se revoit demain mais je suis très preneur. Voilà si vous avez maintenant vous avez vu un peu le premier cours. Comment ça se passe si vous avez des des retours à me faire pour que j'adapte aussi et ça se passe enfin que ça se passe mieux la prochaine fois ?

Alors là t'as 2 policiers optimales qui vont être prendre un côté ou l'autre de leur branchement mais ça dure sur le OK. Par contre oui je crois qu'il est plus court. Ben là le la police optimale ce sera étalée le plus. Du coup j'ai quand même possible de trouver un document qui marche, mais comment est-ce qu'on sait si c'est le chemin optimal ? Bah alors moi le on va le montrer, enfin on va le. Je me satisfais d'avoir trouvé la solution. Ouais complètement je vais je sais pas si on a complètement et ça typiquement c'est bah c'est quoi le paradigme ? Enfin le la notion que tu. Non non non, là c'est un peu plus profond. Là tu dates en train de me dire je peux trouver une première solution et pas chercher les autres. Et du coup ça c'est ça rentre dans le cadre d'une des notions que vous avez présentées hier. Si quelqu'un se rappelle exploration et exploitation, exactement tu peux expliquer pourquoi. Je sais pas qui vient de parler mais dans le cadre de l'exploration on va prendre plus de risques et donc potentiellement causer. Enfin avoir moins de reward pour trouver la solution mais au final ce qu'on veut c'est exploiter. La police. Ouais. Exactement donc dans le cas de mon fait de l'exploitation pure et dur, une fois qu'on a trouvé un chemin, on va pas en chercher d'autre. Mais il faut toujours une exploration. Donc là dans notre cas, comme on a des méthodes ultra brutales, on explore tout l'environnement. Là on a que de l'exploration, donc on sait qu'on sera tous les jours ou mois dans tous les cas. Par contre en effet, si on se place dans un environnement plus compliqué et que typiquement vous avez pas envie de faire évaluation qui va durer, qui va durer 3 jours parce que votre labyrinthe est énorme. Euh et ben vous allez peut être vous dire bah une fois que j'ai une fonction qui est à peu près bonne bah je m'en satisfais et maintenant je fais de l'exploitation et je prends le donc là je prends la policier optimale au vu de la value function c'est pas la vraie donc si on sait que les policiers on prend que déterministes ou pas d'ailleurs on est capable de calculer cette fameuse espérance et donc on connaît notre balle fonction c'est bon pour tout le monde là ce que je suis en train de dire l'espérance l'espérance Après aussi dans notre cas vu qu'on a une politique qui est à priori Stochastique, et Ben on peut écrire la vallée fonctionne comme ça. Donc là ça clarifie le fait que en effet si je trouve la police optimale, bah là la value function optimale associée à la police optimale et Ben là aussi il suffit de la calculer. Donc là on a les sommes. Donc si vous connaissez la politique BAH le fini de Ah sachant si vous le connaissez. Les propriétés de transition les connaissez dans le cas du Mazet sont elles sont binaires, donc en pratique on les écrit pas et ensuite la reward est le la vallée Fonction vous les les connaissez du coup alors lesquels ? Je suis comme ça. Ok. Donc. Pour la police et de l'action policy évaluation, on prend-on prend PI les premiers transition Obama et on en déduit à la fin, quand on a trouvé, on étudie pied. Ouais. Je sais pas, vous avez fait du contrôle ou pas ? Je sais pas si cette slide est très est très. Intéressant si vous avez pas fait de contrôle, c'est pas forcément intéressant. Ouais, ça vous parle pas comment ? Le contrôle c'est c'est une, c'est une matière à part entière de d'optimisation. C'est le pas d'autres de contrôler le fait qu'un agent va. C'est très, c'est quelque chose que tu fais online par exemple si je veux. Je suis pas un exemple de contrôle. T'as, t'as une voiture, tu veux, tu veux ? Tu tu veux optimiser ta tes tes en montée, tu veux optimiser ta vitesse et t'as t'as le contrôle sur uniquement ton accélérateur et là donc le le cadre du contrôle ça permet de dire OK, c'est quoi la fonction ? La fonction appuie sur l'accélérateur qui est la plus optimale pour atteindre l'objectif. Quoi ? Ça se rapproche beaucoup de forcément de learning mais je sais pas s'il faut retenir. Pour ceux qui connaissent pas du tout c'est c'est vraiment quelque chose qui est qui est online. Je fais directement le calcul en entier sur sur. Sur mon enfin sur mon épisode, l'épisode ça devrait vous parler aussi. Et j'optimise directement sur mon épisode alors que là je vais faire l'épisode, je me dis Ah OK là j'ai raté donc je vais recommencer en essayant. C'est pas c'est passé là, mais donc l'idée à qui avait là-dessus, c'est que. On a bien, on va partir d'un point de départ quoi et et avec la trajectoire, avec notre trajectoire, pardon avec la policie on va suivre une trajectoire dans l'espace des des États et on va regarder la value function de enfin la value ou la reward en fonction de notre de notre politique. Donc là ce qu'on peut dire c'est que typiquement notre 2 là. Elle est strictement supérieure à la policière parce que pour n'importe quel patron. Les nappes, donc n'importe quel État dans lequel je me trouve et Ben meilleur une meilleure position que bien in fine il finit au même endroit, mais disons que le P 2 il avait une meilleure chance de gagner que que Pierre. Ok. Et donc toute l'idée de de ce qu'on va faire avec programming c'est calculer successivement. Des, des, des policiers ou des manifestations pour pour converger vers l'optimum. Ça, c'est ce que ça donne si on fait une encore plus d'itération et à la fin, entre autres, ce fameux pied. Ouais OK. Ça c'est le genre de de résultats qu'on pourrait avoir. Typiquement pas dans le cas du maïs, parce que là on voit que c'est continu, mais c'est dans un cas, dans un cas, dans un cas, c'est pas. Ok. Donc ça on le fait de manière itérative OK, donc au début on part d'une, d'une. D'une fonction valeur quoi. Donc là on change, on a plus une trajectoire maison, l'ensemble de nos États possibles OK donc notre ami fonction pour chaque État elle associe une valeur, c'est-à-dire la valeur c'est quoi c'est l'espérance du gain. OK donc on va avoir un premier guest, donc souvent ça va être en fait juste place avec 0 quoi. Et puis petit à petit on va se dire Ah bah en fait dans tel endroit. Dans une meilleure position donc on va avoir des petites bosses qui vont apparaître sur notre sélection si on si on la dessine comme ça. En pratique sur le Mail vous avez pas la destinée, comme ça vous avez la dessiner comme on l'a fait hier, avec des valeurs dans les cases, ce serait plus logique. Donc là, en imaginant que l'État est linéaire, Ben on peut-on peut avoir sur le régime, on va faire des des gestes suivants, et cetera, et cetera, et cetera. Jusqu'à trouver la value function optima, OK. Oui. Tiens, ça, ça représente quoi exactement ? C'est euh cas ici, c'est le l'indice de l'itération qu'on va faire. D'accord, OK on se on, plus on fait de la programmation dynamique, ouais, plus la veille fonctionne, est est est haute, le mieux c'est en gros pas forcément. C'est vrai que là le le dessin est pas. Il montre pas dans les fonctions de. Maximale. Et c'est que est-ce qu'on peut arriver à un cadre ? On a une belle fonction trop grande c'est pas sûr. Enfin sauf si si on peut en itérant avec une value fonction trop grande dès le début donc si c'est possible. Mais l'idée c'est que on est dans un processus d'édition Markovien donc en MDP donc en fait cette value. Elle existe. De enfin, associé à la policier optimale, elle existe. Donc en fait on n'a pas le y a pas de notion d'une value function qui est meilleure que notre, c'est juste que c'est la vraie quoi. C'est la vraie value function ? Ok en en enfin le lever lever star il est unique quoi, il existe et tu pourras pas, tu pourras en trouver un par contre. En termes l'estimateur, oui là vesta, c'est le meilleur des estimateurs qu'on puisse avoir de vesta et logique. Ok OK, donc il faut nous par exemple, on peut graviter autour à chaque fois en en jouant avec. Avec. Avec ouais Bah on pas on va vu qu'on la connaît pas au début. Typiquement un cas classique ça va être de l'initialiser à 0. On va faire des trucs avec ça. On va découvrir des choses sur notre, sur notre environnement en explorant l'explorant donc on va avoir plus d'informations donc on va. Je vais pas dessiner la petite bosse qui est qui est là au début, vous pouvez pas voir là où je montre. L'intérieur et regarde le mur, c'est trop bizarre, enfin bon. Mais donc petit à petit vous allez avoir plus d'informations. Donc votre vraie value function que vous estimez elle va se rapprocher de la valeur réelle ? Ok c'est tout le but de de désagrément qu'on va implémenter justement. Monsieur, donc y a pas de souci. Maintenant, on commence à rentrer dans concrètement ce qu'on va implémenter. Quelqu'un peut me dire ce que ce qui a marqué la la première formule au dessus ? Je te demande pas de me la lire M. Très bien, quelqu'un peut faire plus. Non, très bien. Vous reconnaissez ? Vous reconnaissez quelque chose qu'on a vu quand même, non ? Si si j'enlève les si j'enlève les cartes. Ouais, exactement, c'est juste l'équation de Bellman. C'est clair pour tout le monde que c'est une belle mal là si j'enlève les cas. C'est-à-dire la forme avec l'espérance et la forme avec la somme, c'est ça ? Bah la définition. Enfin la, la version la plus compacte de bellman est la plus générale, ça avec l'espérance. Mais en l'occurrence, vu qu'on est en cas discret, tu la développes de cette manière là ? Mais typiquement, si on avait un cas qui était continu. Enfin je c'est pas bien de dire ça parce que je parle pas de la lettre K, je parle du de l'environnement quoi. Si on avait un environnement continu, on pourrait pas faire la somme comme ça on pourrait faire une intégrale donc on l'écrirait. On écrit cette formule là sous forme intégrale. OK. Donc donc c'est en effet c'est la même formule. Simplement là on la développe parce que on sait que derrière on va vouloir la programmer et donc on programme pas une espérance. Encore une fois on relance. Ok du coup là ce qu'on fait ici c'est juste vous prenez bellman du cours précédent que vous avez bien noté que dans la fiche et que je vous ai dit beaucoup de. Une fois de de bien connaître. Vous prenez bellman et vous vous dites OK. Si je suppose que ma value function allait à K ? Elle est, elle est, elle est vraie, elle est pas trop loin alors normalement. Enfin si elle est optimale, alors je suis censé vérifier belman et donc. Et donc y a +1 il est censé enfin enfin pardon et donc si je prends l'état S et je l'explique comme ça je vais pouvoir le lier grâce à ça à mes États +1 en pratique ce sera pas vrai parce que votre votre fonction votre relation elle est pas elle est pas exacte. Par contre vu que enfin si vous forcez disons la la variable BK +1 enfin la la value fonction K +1 en SA être égal à à ça. Alors vous allez probablement vous rapprocher de la réalité vu que vous forcez votre ville fonction à. À pardon, à à vérifier. Elman. Et donc du coup ça vous donne une manière d'itérer sur votre value function ? C'est très important, là on se rend compte que grâce à belman si j'ai un premier guess ? Donc une première estimation de V alors je suis capable d'avoir une estimation avec en +1. Donc là normalement je peux vous lâcher sur Python et vous dire de le de l'implanter. Vas-y, je t'en prie. Pas mal de questions, en fait, me perturbe. C'est l'innovation. Alors merci quoi ? RAS. Ah Yes Yes. Oui c'est c'est c'est pour l'indice. Donc j'avais déjà équipé comme ça la dernière fois. Là c'est juste que il y a. Les toutes les nations sont pas toujours homogènes. Ça c'est c'est ma faute. Je devrais ce que je récupère des trucs d'endroits différents, mais. Mais c'est bien posé la question si jamais c'était pas clair pour d'autres vous avez 1R, vous savez que c'est une Ward AIS parce que vous l'indiquez. Donc ça dépend de AIS. Ok. Euh et la priorité de tradition ici, c'est bon ? Oui B d'aller en exprimer sachant que en s en première action AOK dans le cas du Maze ça existe pas. Euh. Donc est-ce que c'est clair pour tout le monde que là on peut implémenter la value éthérisation déjà ? Non ? On avait, on connaissait s ouais, en supposant que tu connais avec. Parce que on a une première estimation de l'attraction de la vallée. Déterminez. Est-ce en sachant que est-ce que vous en. Oui, tout à fait, et donc comment est-ce que là on peut le connaître ? Bien en allant, c'est. Moi attends c'est c'est c'est bien. Si vous reprenez le graphe en haut, c'est ce que vous avez hier. Là on le faisait dans le temps. Ça veut dire que si si je connais les value function qui suivent mon état, là je peux en déduire la value function qui sont à mon État maintenant. En effet, dans un cas un peu différent. Je connais veka partout. Ok enfin je le connais, je ou je le définis pas. Par exemple on prendra souvent 0 au début ou ou des valeurs Random. Donc je le connais, je connais mon VK. Par contre, je sais que mon VCA, il doit vérifier bellman s'il est optimal. Une hypothèse que on est on a déjà, on a déjà trouvé, on a déjà trouvé un et on fait une 2nde estimation en se disant, Bon, maintenant je vais essayer de raffiner mon Vega pour qu'il soit plus proche, de vérifier bellman. Et donc en fait c'est comme si dans mon graphe de de d'hier hein j'enlève une une note. Enfin j'enlève un nœud exactement comme les petits de ce qu'on a fait. Et après je demande bah recalculez moi la valeur en fonction des valeurs qui a autour. Donc si tu supposes que tes valeurs autour sont pas trop mauvaises, alors Ben on peut. Alors tu trouves la l'est encore meilleure. À l'heure pour la note que tu viens donc les là c'est ce qu'on fait, on enlève la, enfin on on se place en en s ouais hein, et grâce à l'information qu'on a autour de s, on propage sur notre État s donc ça ça me fait un une nouvelle estimation de s parce que enfin V 2S, on l'avait, on l'avait déjà, on avait une valeur pour V 2S. Donc simplement là, en utilisant bellman, on a une nouvelle valeur de VK de s, donc on l'appelle VK +1 de S. Parce que cette nouvelle estimation de l'opération, c'est une nouvelle, une nouvelle opération. Ouais, c'est ça on est, on est dans dans ce cas-là, on a une première estimation. On veut calculer, on se met au milieu à avoir pour un État particulier, on veut que la nouvelle valeur de V. Plus proche de la réalité ? Donc à à l'itération K et on utilise pour ça les voisins de la de l'État RS pour calculer pour quelqu'un l'heure OK. J'ai ouais ouais oui bah après ça on l'a et même y a des y a des exemples plus tard, mais ce que l'idée générale est moins comprise, pourquoi est-ce qu'on fait ça ? Et pourquoi est ce que la la, la fonction qui est là, vous pouvez déjà l'implémenter en soi parce que vous avez toutes les informations. Bon bah. Ok. Assis. Quelqu'un peut parler ? Vous chantez Léonard suppose que. OK, ça devrait donc OK donc donc toc toc. Ouais donc là maintenant on se remet dans le cas du du labyrinthe. Donc comme je vous le disais on va représenter notre grille pour pour le bonus function et dans chaque case on va vous, on va pouvoir mettre la valeur qu'on va trouver. Donc là on est dans un cas très simple où le. On on part de d'en haut à gauche et on veut descendre tout en bas, tout en bas à droite. Mais t'étais une. Et ouais, et donc là. Notre récompense c'est pareil qu'hier, à chaque fois qu'on fait une itération on on prend moins un et en gros faut plus de zone le plus vite possible sortir du labyrinthe et pour l'instant on suppose que notre agent il se déplace de manière Random sur les OK c'est bon pour tout le monde. Oui, c'est bon. Hop. Maintenant là, je vous donne un exemple pour. Je vous donne un exemple de ce qui va se passer si on applique ce qu'on vient de ce qu'on vient de dire, OK. Donc là on est, on est pas sûr de la. Imagine on va, on va attirer sur la policier directement. OK ? Donc là la première politique. C'est la politique Random, OK, donc. Pour la politique Random sachant c'est pas. On sait pas ce qu'on fait, elle est, on sait notre, c'est notre notre premier qu'on simplement l'initialise comme ça. Ensuite ? Si on fait une étape en dessous là un, on remarque que. Pour la police sérandon, si on est juste à côté de. De la fin de de l'arrivée alors ? Et Ah oui pardon je me suis-je me suis en me fermé les pinceaux. On a 2 étapes terminaux c'est pas un début et fin, simplement on gagne si on arrive dans un des coins à partir de dans notre dans notre coin. Donc en pratique on pourra faire commencer notre argent n'importe où une c'est qu'il a il le plus vite possible ? Pour un des points de vue. Voilà donc on fait une évaluation des nôtres policiers en on fait un pas un peu n'importe où et donc là. À notre mère itération, on va avoir moins un partout parce que on va, on va récupérer nos récompenses de moins un parce qu'on a fait une étape dans le dans le labyrinthe et une récompense de 0. Et vous allez voir que on obtient un petit à petit. Bah en fait si on sait que notre value function c'est bien que de moins un juste à côté et bah et que c'est 0. Et surtout que la vallée Fonction c'est 0 à l'endroit de la case. Bah en fait les États qui sont juste à côté on va être capables en va être capable en en ayant. En ayant ensuite une politique qui est plus Random mais qui est qui est gris, ça ça devrait vous parler aussi. Donc c'est juste. On optimise un maximum afin de prendre toujours le chemin optimal. Et Ben là on a la fameuse propagation dont je vous parlais ou on on propage nos étapes petit à petit. Et on. Et petit à petit, on va converger vers la politique optimale. Donc là en fait, dès qu'un égal 3 on a la politique optimale, AK égale un. Il y a uniquement les cases qui sont juste à côté qui peuvent comprendre que l'arrivée est à côté parce qu'elles ont accès à l'information. C'est pareil mais pour 2 cases à côté mais mais les les la dernière diagonale qui est à 3 cases de l'arrivée elle a pas encore l'information. Et à partir duquel il y a 3, on a déjà la la police optimale. Par contre vous allez vous allez vérifier que. Que en fait la ? La value function elle a pas encore convergé, ça veut dire que on peut avoir trouvé la police optimale, c'est-à-dire que la value function open donnée suffisamment d'informations. Dès. Comment dire ? La value fonctionne peut être suffisamment proche de la réalité pour que ça suffise pour trouver la police optimale, je sais pas si c'est très clair. Ici on a notre religion qui a certaines valeurs qui sont pas exactement les valeurs qu'on attendrait. Mais. En en prenant ces valeurs là donc en prenant la la la politique gloutonne. Voilà avec des policiers. Bah en fait c'est suffisant. Parce que le tout ce qu'on veut en fait, c'est que la diagonale soit le chiffre le plus grand, la 2e diagonale un peu petit et la diagonale encore le plus petit. Parce que on sait qu'on veut aller vers les les les plus grandes, les plus grandes choses. Par contre, c'est pas encore la vraie valeur de la vallée d'action pour cette policière. Exactement là on avait ça là on avait star là on a pu star là on est pas star mais on n'a pas encore vu star ok c'est un peu comme si si on revient d'ici là en fait ce qui compte c'est que notre agent est sage verrou aller donc en fait il compte c'est d'avoir toutes les aspérités les aspérités Elle est elle est déjà bonne en fait, elle est optimale. Par contre j'ai pas encore vu ça parce que c'est pas la vraie valeur de. Il y a. Est-ce qu'il y a encore d'accord ? Là on est à l'accueil 3 et il y a 3 nouvelles dessus. Il y a une dimension, un problème de savoir milité, ration. Il se termine le non, l'itération de se faire ? Bah là en l'occurrence pour le labyrinthe on pourrait arriver rapidement à montrer que le nombre d'itérations. C'est pour avoir la police optimale. C'est c'est quoi quelqu'un, une intuition là-dessus ? Ça, c'est pas en carré. Parce que là on a gagné 3 donc c'est pas c'est pas vrai. Ouais ouais. Je pense que vous pourrez enfin quelqu'un peut trop c'est pas. Là, on a vu que on propageait depuis les terminaux. Ah c'est le nombre de Pâques pour attendre un minute maximum de Pâques si on. Si on est, si on. Les acteurs ne pas pour arriver. Ouais, vous avez entendu sur sur teams. Oui, OK donc. Et ce que vous êtes d'accord, ce que vous avez compris, ce qui vient de dire. Oui. Aussi, il a dit, C'est que la, le nombre d'itérations nécessaires pour. Pour résoudre le problème ici, pas pour trouver la bonne fonction, mais pour trouver la la. La Poly optimale c'est égal sachant que on a une reward de moins un et que en gros, à chaque pas de temps on. On propage d'une fonction de un, donc on a plus d'informations au niveau de un de plus et donc dans le cadre du labyrinthe, il y a un cadre très simple en fait, il suffit que on prenne le chemin le plus long et qu'on puisse propager la jeune jusqu'à ce chemin. Le plus. Parce que siège parlent forcément, tous les autres chemins sont plus courts, donc ont déjà eu la la propagation. Voilà donc là ça vous donne le premier algorithme que vous allez pouvoir implémenter tout à l'heure qui est l'itératif police évaluation. Et donc là, enfin, c'est vraiment exactement ce que ce que je disais. On va itérativement évaluer notre politique. Monsieur oui j'ai une question s'il vous plaît, j'ai pas compris ce que vous avez dit sur la slide précédente. À propos de fois ce que Léo avait dit là pour ce que Léonard avait précisé, pour le le fait de prendre le chemin le plus long et de propager, j'avais pas trop compris de quoi c'est une alors attends je vais, je vais sortir d'autres slides. Euh hop. Voilà donc là c'est sur, c'est sur la value itéré chaîne ici. Ok on va, on va juste après, mais donc dans l'idée au début on a 0 partout et on sait qu'on a un parce que. Parce que on est arrivé, enfin, on est sur les terminales. Et petit à petit, à chaque itération propage l'information qu'on a. Avec. Avec une nouvelle libération. Et donc ce que dit Léonard, c'est que enfin, la question de Léonard, c'était. Pour l'intégration, il faut. Donc sachant que dans ce cas-là très simple, on propage de un à chaque fois, et Ben le nombre d'itérations dont on va avoir besoin pour résoudre, pour résoudre, pour avoir la police optimale, c'est égal au nombre de enfin à la distance maximale que l'on peut faire dans le labyrinthe. C'est sûr que si je place mon mon point vraiment tout à la fin. J'aurais besoin que d'une itération, mais si je le place tout en bas à gauche ici, Eh Ben je vais avoir besoin de de du chemin maximal. Donc j'ai besoin de faire toutes les itérations pour arriver à la vallée, fonction de terminer. C'est plus clair comme ça. Ouais, ça c'est mieux. Merci beaucoup. Voilà on ça en vrai, on le reverra tout à l'heure. On en tt ce sera plus simple de le voir en vrai. Il est 08h54. Oui ouais, TAC TAC TAC ici. Ouais Ouais, on va revenir là-dessus, après je préfère faire la value. Ch. Point ouais on va faire la la value direction tout de suite. Je reviendrai sur la pollicitation après. Hop donc là ça ressemble plus à ce dont je parlais directement, je pense que c'est plus logique de commencer comme ça. Déjà présenté ? Ouais, Ouais, OK. Alors ici ? On on va faire comme on a fait tout à l'heure, on va pouvoir. On va comment dire ? D'une autre manière de de voir le problème, juste avant, on faisait, on utilisait de l'allemand un peu tel quel. Souvent rappelé, on est comme ça, on l'utilisait comme ça, OK. Sauf que là, maintenant. On va supposer que on. On a la politique optimale, donc on va là, on là, on évalue notre policier. Ok, on a une politique qui est fixée et. Notre politique là, puis elle est, elle est fixée, et simplement, on l'évalue, c'est-à-dire que une fois qu'on a notre notre policier. On trouve la relation qui est associée OK et et ça, ça nous permet à posteriori de. De comment dire ? Une fois qu'on a la value function on va pouvoir améliorer notre police. Ok ça permettra de de de de faire la fameuse de police interrogation. Maintenant, on fait un truc différent. Alors hop. Hop hop hop hop hop ouais ouais. Ça, c'est pour la sécher, la sécher. Alors ? Maintenant à la place de de d'estimer directement ma value fonction, je vais estimer ma politique. Je vais quand même utiliser bellman parce qu'on a vu que c'était à peu près à peu près équivalent. Simplement je vais changer la formule de tout à l'heure. Donc là vous voyez ce qui est implémenté au niveau du bellman. Il y a un Max qui a apparu. Est-ce que vous pouvez m'expliquer pourquoi ? Pourquoi est-ce qu'on a un Max maintenant, sachant que on suppose que on a trouvé une politique optimale ? Ça ? Tu peux expliquer un peu plus en détail pour les autres c'est certain. De savoir exactement, donc là on est. Ici, on était avec une politique qui. Qui déterminé ? Et on va, on va avoir les les probabilités de de passage maintenant. On sait que. Notre politique est optimale. Et on suppose qu'on a la validation, donc en fait pour avoir. Enfin, sachant que ma politique optimale il suffit de prendre le Max de la value function en fonction des États, donc pour toutes les États, enfin pour les pardon pour toutes les actions possibles, je prends la l'action qui me donne la la qui me mène vers la meilleure relation. Donc si ou alors je peux le voir en termes. De que function ou je prends le la valeur de Q telle que enfin la valeur maximale de Q en fonction de a bien strictement même on a on a vu hein. C'est soit je dis je prends, je prends l'action. Ah donc je regarde l'État qui est derrière. Soit je prends directement la valeur maximale de Q, sachant que Q lit à la fois l'État et la. L'astuce pour rajoute, c'est que on prend en compte là-dessus aussi exactement. Donc les valeurs sont différentes, c'est juste une manière différente d'écrire. Décrire un problème de parentalité quand c'est du REM, concrètement, Ouais, ça prend toujours en compte l'action et jamais dans le cas d'une fonction. Alors ? Oui, on prend toujours en compte l'action. Après t'as les fonction en elle même. Suffit pour choisir tes actions d'air. Parce que si tu sais dans quel État tu veux arriver. Est-ce que tu sais dans quel État tu veux arriver ? Il suffit de prendre l'action qui permet d'aller dans cet État là et ça, ça implique en effet de savoir quelle est la, enfin de savoir quelle action faire pour arriver dans l'état de la famille. Donc c'est pas c'est pas évident. Dans le cas du du labyrinthe, c'est évident. Si vous savez vous voulez aller à la cage 110 et vous êtes en 0 0, Vous savez quelle action faible. Si vous êtes dans un jeu de société. Enfin dans le jeu de société, c'est un jeu de table cible facile. Mais si enfin un robot il sait dans quelle position il veut se trouver, c'est pas évident d'avoir de connaître l'action à faire pour arriver dans cette position. Donc en fait, une fois qu'on a toutes les actions possibles et qu'on sait quelles sont les bières, on peut s'en passer parce que c'est des exemples. C'est pour ça que, enfin c'est pour ça que je fais beaucoup dit hier que si j'ai la police optimale ou la cuve, la cuve table optimale, alors en fait on a déjà la position. Ok, OK. Alors ? Toc. Ouais donc ça c'est notre algorithme qui permet de. Euh. D'évaluer notre policier ? Ok, c'est ce qu'on viendra juste avant. On a notre policier. Qui est fixé et. Et on calcule grâce à ça. La Belle. Je sais pas ce que c'est que la brute est claire ce que je vous ai quand même bien expliqué la la la formule ici on l'a dit, on l'a déjà bien détaillé à chaque fois on régime notre value function jusqu'à ce que notre la différence entre la value function à 4 et K +1 donc notre prochaine itération. Elle soit suffisamment faible pour qu'on puisse se dire qu'on a convergé. Ouais. Okay ? Tu parles de cette formule là ? Ouais mais le terme de couche ? Le premier terme ouais on est sur de l'évaluation tu vois en premier match de d'épisode chrome policier optimal pour reprendre un policier optimal pour déterminer la value fonction optimale. Ça ce qu'on va faire juste après. Mais là pour l'instant on fait un truc encore plus simple, c'est juste. C'est juste euh non, juste laisser le policier à la police optimale combattant. Et là on voit qu'on peut le faire parce qu'on l'a, on la trouve à causer. Par contre là ce qu'on trouve ici et le et le le le le but de la police évaluation c'est de trouver la value fonction associée à la policier psychopathe. D'accord ? Ah pardon à la policier. Quelconque. Là on trouve vpi tout court. Y a pas c'est pas, c'est pas vpi star c'est vpi tout court. Oui OK par contre à droite ce qu'on remarque mais c'est on le remarque mais on le. Comment dire ? On n'a pas cherché à le calculer pour l'instant, mais par exemple si c'était un calendrier, ouais je vois que star je dois refaire ce que j'ai fait là avec ces policiers là. Donc si je veux avoir la le le veste a associé oui mais en l'occurrence enfin on on va voir mais on comme comme ça un peu particulier ce qu'on est dans le car Random. Donc dans le cas Random en fait on sait qu'on explore tout et donc en général le cas Random nous permet de trouver la valeur optimale. Donc là pardon de trop de retrouver la. Pardon la la value function associée à la policière Random OK qui me donne l'espérance de mon gain sachant que je fais un truc Random quoi. Il me donne quand même accès à suffisamment d'informations pour en extraire la police optimale. Ok OK. Policier, un homme pense être l'homme, une greffe si c'est par exemple. Ouais dans dans des cas, dans des cas compliqués, dans le cas de labyrinthe la Random suffit largement. Parce que du coup comme on prend la Random policier, on sait que on explore tout. Non en vrai je n'aurais pas dû chercher à vous montrer la littérature tout de suite en fait en mais donc du coup ça c'est un algorithme qui est le premier que vous trouvez, qui est donc le plus simple qu'on a. Pour l'instant ça ne permet si je vous donne PI. Une \*\*\*\* c'est. Pis la politique qu'on évalue si je vous donne PI ? Avec cette méthode là vous me calculez vpy associé donc la fonction qui a à chaque État associe l'espérance du gain sur l'épisode. Un Monsieur, j'ai une question, ouais. Je suis la slide précédente. Ouais euh là la 2e sonde elle est contenue à la première somme ou alors c'est le produit de 2 sommes. C'est connu, c'est pas le compte. Ok parce que tu vois, tu tu étires d'abord sur les actions. Et ensuite sur les États sachant l'action. C'est donc ça c'est l'indice qui doit te faire dire que tu sommes bien sur les 2 en même temps ? Je sais pas si c'est clair, ouais. Et Monsieur juge, une question, voilà là l'algo qu'on a vu, c'est trouver la enfin la la value fonction optimale qui nous permettra de trouver la politique optimale en partant d'une politique aléatoire. Attends, refais la OK. Le but de cet algorithme c'est de trouver la politique afin de trouver la value fonction star qui nous permet de trouver la politique STAR en partant de la politique aléatoire. C'est ça, mais juste attention, on va trouver. On rencontre pista en effet. Mais on n'a pas calculé westar comme j'ai dit. On a calculé et puis Random. Et puis ouais, voilà, c'est ça. OK, ça marche, merci beaucoup. Et c'est pas testage devant ce que vesta, on sait ce que c'est. Les stars on sait que c'est dans le dans le dans le cas très simple hein. Les stars ça devrait être ici si je me mets tout en bas là. L'infini ? Si je prends la première liste, ça devrait être 0-1-2-3, on est d'accord ? Ça c'est ça. OK, c'est bon navire. À chaque fois on recule de, on recule de moins un parce qu'on sait qu'on prend le chemin optimal à chaque fois. Là ce qui est ce que vous pouvez remarquer, c'est que l'espérance de. Mon gars si je suis juste à côté de de l'arrivée c'est pas moins un parce que ma police est complètement Random donc j'ai quand même si je me place taro j'ai encore 2 chances sur 3 de partir dans la mauvaise direction. Ce que je veux dire ? Par contre, j'ai quand même plus d'espérance de de faire bientôt parce que j'ai quand même une chance sur 3 quand je me place juste à côté dans les coins. Que de manière Random arrivée sur l'arrivée. Ok, et c'est pour ça en fait que on c'est. C'est c'est pour ça qu'on peut avoir l'intuition que grâce à cette value function on peut trouver pis star parce qu'en fait nous donne quand même suffisamment d'informations pour nous dire en fait si tu explores aléatoirement, Ben en fait la vallée fonction sera quand même plus plus faible, enfin plus grande. Pardon proche de l'arrivée et donc définit si ensuite je passe sur une crédit politique et elle les sert juste à optimiser. À la value function. Euh Ben ça suffit et c'est ce que c'est ce qu'on avait juste au-dessus. Là c'est l'idée que en soit la la politique. Édit, enfin glouton, pardon elle fait que elle fait que la de la montée de gradients sur la fin même pas de montée de gradient, juste elle va chercher à remonter la belle fonction. Si jamais ma value function elle a déjà toutes les aspérités. Ils sont nécessaires dans. Enfin comment dire si elle elle présente déjà toutes les toute la géométrie nécessaire. En fait, ça suffit pour avoir la police optimale. Ouais. Ouais. Les courbes de lever, en fait, on pourrait trouver. Par exemple, enfin, dans le sens des variations de la Cour, ça c'est exactement les mêmes, mais pas sur le même value, on est déjà sur la police optimale. La policier en fait, c'est la variation de la peau. La forme est exactement la policier, là ça va juste pas la policy l'automne, attention hein, ça va être ok. Je regarde là, c'est le plus grand, je taxé de moi, mais une fois que j'ai terminé votre déception mal. J'ai une forme de cœur particulière. Ouais, pas pour la. Pour la policier oui mais du coup là on visualise la vallée Fonction valise, pas la la la la. Les variations de la course, c'est la. La politique dit Ouais. Parce que typiquement si je passe une politique et stochastique OK Ouais OK ce qui est important c'est que le la la politique de l'automne elle elle elle elle cherche pas à regarder des valeurs qui sont différentes et regarde juste OK c'est quoi la valeur maximale que je trouve juste à côté ? Donc c'est pour ça que il suffit d'avoir les les variations même très faibles de V et que je peux mettre V divisé par 10 ce sera pareil. Par contre si je passe sur une politique qui est stochastique ou j'imagine que la probabilité d'aller dans la direction dépend de la valeur justement ? Bah là si je divise par 10 en pratique je vais beaucoup. J'avais beaucoup plus ressembler à une une politique qui est Random je comprends. Ouais vraiment la question c'est sur la ouais. Sert à. Pour les États ? Les États c'est toutes les cases, c'est ça c'est vrai. Sauf qu'en pratique, la probabilité de sauter 2 cases, la probabilité de passer la case de 0 2 sachant que t'es en 0 0, elle est de combien ? Non, elle est nulle. C'est si si la propriété de sauter de case elle est nulle, c'est impossible de lire. C'est pour ça que je vous disais la probabilité. Là elle est définie sur toutes les toutes les pertes d'espace. Mais en pratique, c'est une probabilité qui se parce enfin qui est-il y a beaucoup de zéros dedans. Donc en pratique on ici on va beaucoup écrire enfin au niveau de la somme sur les États pour tout exprime dans les voisins de l'i s tu vois ? Parce qu'en fait on sait que tout le reste c'est 0 donc on va pas calculer la politique et même en pratique. La probabilité de passer en s prime RSA, ce sera un dans notre cas dans le cas du labyrinthe. Pourquoi ? On en a parlé hier si je sais que c'est pas évident mais le le MDP c'est un c'est un cas très général, on a des probabilités de transition sachant et actions. Sauf que dans le cadre du labyrinthe, si je prends la décision d'aller. À droite ? Personne va m'empêcher donc la probabilité. Donc là, si je le réécris, la probabilité d'aller en 0 un sachant que je suis en 0 0 et que je veux aller à droite, c'est un. Yes mais y a plein de c'est quoi exactement ? Donc en pratique là dans le dans le labyrinthe. Donc là bien entendu vous avez un un algorithme qui est dans le cas général, donc c'est plus difficile à appréhender parce que en effet il y a une probabilité de transition et il y a il y a des États, y a une politique qui est stochastique en pratique quand on va implémenter nous sur un cas très simple. Typiquement la probabilité de transition elle existe pas parce que elle est de un, donc on va pas les faire. Et en pratique on va pas écrire la somme sur tous les États parce qu'on sait que tous les États ne sont pas liés entre eux. Et on va écrire la somme sur les États voisins de SOK, ce qui est strictement équivalent à cette cette formule là elle reste vraie. Simplement là ici ce pays sera tout le temps égal à un. Ici, la somme, elle sera réduite à juste les voisins de notre case et. Et notre paix là, ici, ce sera constant. Si c'est le cas de la grippe 2, la Renault Poly. Il a levé la jambe pour faire une marche. Ouais la poêle transition elle est pas devant parce qu'il doit y avoir un obstacle qui l'empêche et étant c'est ça ou oui un problème technique ? Ouais Ah ça dans la vraie vie tout est stochastique donc il faut-il faut se placer dans un cadre qui nous permettent de modéliser ces orages. On peut vouloir faire une action mais il peut se passer un truc. Et donc juste on est. On est content parce que nous on a un cas général et simplement dans dans nos canaux il va se simplifier donc ça va être plus sympa à. Qu'est-ce que c'est que pour tout le monde de la police ? Évaluation. Oui ouais bah 10 secondes pour poser une question alors je passe à la suite. Ouais, il y a encore, il y a encore 4 slides, on fait une pause après. Ok. Bon, maintenant on a un algorithme qui nous permet de. On a un algorithme qui nous permet de trouver la fonction associée à une police. Comme on a vu. Juste au-dessus là. On s'est dit Ah mais en fait maintenant que j'ai la value function associée ? À partir de ces valeurs fonction, en fait, je peux. Trouver la police optimale. Donc. Ça veut dire que là, j'ai trouvé une première méthode algorithmique de trouver une nouvelle politique qui sera bonne pour mon agent. Ok, donc ça, ça va s'appeler là la policy atterrage. Ok donc la police va lui d'abord. On on prend une, on part d'une politique qui rentre d'homme, OK ? On parle d'une politique qui Random et. On va ? L'évaluer, on va trouver la la value function associée grâce à l'algorithme qu'on vient d'implémenter. Et une fois qu'on a l'évolution, on va se dire on va très bien, maintenant on va faire la, on va prendre une nouvelle politique et la politique glouton associée à la value function qu'on vient de trouver. Est-ce que c'est clair ou pas ? On vient, on vient de voir que on a trouvé une politique qui serait une bonne candidate, donc il suffit de dire maintenant. Et et et on en extrait la policier optimale. Donc en pratique ça serait comme ça, c'est la nouvelle policy Pie prime. C'est enfin à s c'est l'Arma Max de Vpi de s pour toutes les actions possi. Ouais donc pour toutes les pour tous les États voisins possibles. Ok, donc on prend notre policier et notre belle fonction, on en extrait à la policy et on une nouvelle policier. Donc en pratique, ça on va le faire plusieurs fois, parce que peut-être que pour la première itération, le rendu c'était pas suffisant, qu'on avait des premières, des premières variations, mais c'était pas suffisant. Donc on prend notre polyanthum, on l'évalue, on en extrait une prochaine policier qu'on réévalue, on a une nouvelle fonction et on recommence ça. Chaque correspond à un épisode ou c'est pas tout ça, un épisode c'est. Ouais là ça va être différent, du coup on va parce que là en fait on est en train de faire du tabulaire, on fait pour tous les États à la fois, comme on a vu notre notre algorithme là. On update notre V 2S pour chaque s qu'on connaît. Donc c'est un peu comme si tu explorais tout, tout, tout, tout, tout et tu tu regardais ça. C'est épisode d'une certaine manière, ouais. Après là où là où je suis pas très fan d'utiliser le mot épisode ici, c'est que souvent en R un épisode c'est vraiment je lâche mon argent dans dans la jungle, je suis une trajectoire. D'accord, OK, alors que là on prend toute la grille et on regarde un par un les États auxquels si je viens en l'État 0 0, c'est quoi ? OK, si je suis en état un un, il se passe quoi, et cetera. Donc je suis pas vraiment en train de faire un épisode dans le sens je suis pas en train de jouer au jeu, je suis en train de jouer au labyrinthe quoi, là je suis en train de juste tout bousiller. Je fais OK, si je me mettais dans un labyrinthe, il se passe quoi ? Si je mets là, il se passe quoi ? Si jamais il se passe quoi ? Je retiens tout et ensuite je dis OK bah du coup je sais tout et donc je peux en déterminer. Donc itérative policier, interrogation, on a pierro qui en général la policière Random qui me permet toute explorer, on en extrait vpi 0. On en extrait une nouvelle politique, a et cetera, et cetera, et cetera, et on sait, on peut le montrer, que à la fin, ça, ça converge vers pista et Vista. Hop. Donc là c'est des petits diagrammes pour ceux qui qui comprennent mieux comme ça, on part de PI, on évalue, on AV et donc on a l'évaluation et les improvement. L'amélioration OK voilà, c'est bien. Il y a plusieurs diagrammes, donc vous voulez garder en tête celui qui vous parle le plus. Mais mais voilà, et on sait qu'à chaque fois on va converger vers les 2 qui sont. Donc uniquement ça donne quoi ? On prend un bloc de policiers, évaluation qu'on avait déjà. Et ensuite on fait le policier proprement, donc. Quelqu'un peut expliquer le CE qui se passe en 3 hein ? Comment est-ce qu'on en 3 c'est la police qu'on va ? C'est-à-dire qu'on extrait la meilleure politique à partir de la value fonction qu'on a déjà. Donc comment ça marche ? Pas on va, on va, on va trouver la chaîne qui est associée. Et. Ensuite ? Et Ben on va. Exactement. Alors pourquoi est ce qu'on ? Pourquoi ce qu'on par tous les États ? C'est évident ce qu'on veut trouver pour tous les États. Quelle est l'action qu'on va faire ? OK, donc à partir du moment où on est dans un État, qu'est-ce qui se passe ? Qu'est-ce qu'on prend pour PI de s ? Pourquoi ce qu'on a un arc Max là ? Pour essayer de trouver de la. Pourtant. Ouais c'est ça. Et donc tu peux m'expliquer c'était quoi dans l'arnaque ? C'est pourquoi est ce que on utilise ça en armax ? Je crois que. Marche, c'est une des valeurs associées. Maximum possible. De du coup, notre. Ouais bon c'est l'idée en gros c'est que. Euh, notre policy, si elle est optimale, alors elle maximise la vallée fonction, c'est-à-dire que. Si je suis, si je suis dans un État, si on retourne là-haut, là. Comment est-ce qu'on extrait la police optimale à partir de la fonction ? C'est simple, on se place dans une case et on regarde dans mes voisins. Où est-ce que se trouve la vallée, fonction la plus grande ? D'accord, on se dirige vers les, on se dirige vers les endroits où la vallée fonction la plus grande possible et ça ça me donne la direction que je dois trouver. Donc là par exemple y a 2 flèches parce qu'en fait il y a 2 voisins qui ont la même value function donc je pourrais faire les 2, c'est pas grave OK ? Donc hop. Donc in fine, ça veut dire. Ici, on évalue. Notre vallée function. Avec les primes ? Ok. Et d'ailleurs on pourrait aussi complètement utiliser V de St Direct. Ah là du coup on REC avec notre solution qui mène tous les États dossiers. Bah c'est ça en fait. Pour chaque nouvel État, on va retenir. OK ma policier, c'est que c'est encore une fois on reste sur notre tabulaires, donc une policier, là, dans votre tête, la policier sur un labyrinthe, c'est une grille avec des flèches sur la grille. Ouais enfin notice avec pour chaque endroit, pour chaque endroit, je me trouve dans le labyrinthe, verrou, je vais, c'est pas, on n'est pas encore sûr, nouvelle compliqué où j'ai une fonction qui va me dire en fonction de mes entrées, je vais faire mes sorties. Là on retient tous les cas possibles, on fait, on fait, on dit, on fait du tabulaire. Donc on traverse tous les États qu'on connaît pour tous les États qu'on connaît. On on se dit OK bah c'est c'est c'est vers où que je vais pour aller pour avoir ma valise fonction ? Donc on prend leur max. De et l'histoire de Polycristal, c'est juste pour vérifier que que. Que ma police ne change pas. Ok, donc ça veut dire en gros pour chaque nouvelle, pour chaque nouvelle nouveau changement. Pardon que je fais. Donc si old action est différent, rapide, c'est-à-dire si. Si j'ai changé ma policier, alors ça veut dire qu'il y a encore des des des des changements à faire sur ma policier. Donc ma police est pas encore stable. Donc je veux dire que je vais évaluer ma policier réévalue ma fonction. Et ensuite, je vais revérifier une dernière. Bref donc en fait ils ont dit tout à l'heure entre nous sur les grilles avec une droite. Donc c'est tout à l'heure, une fois ici à droite. En fait, on a été ici très stable, on repart de début pour évaluer la vallée, fonction en fonction des races anciennes rares. Donc là dans ce cas-là il y a besoin que d'une seule libération. Parce que enfin l'important on va avoir 2 itérations ici. Parce que on va avoir une première itération où j'ai calculé ma value function. Ouais. J'en extrais la politique de la police optimale. OK là elle est optimale mais on le sait pas encore et du coup une fois j'ai ma police optimale je vais réévaluer la police version et là au bout de 2 cas égal 3 j'ai directement ma valeur parce que parce que on a la politique optimale et donc j'aurai ma ma V 2. Qui en réalité ? Vesta. Mais on sait pas encore. Je réévalue ma policier et il dit, Ah bah c'est bon, là j'ai la même, j'ai la même valeur et la même. Enfin, pour pour une policier, j'obtiens la même la même vue fiction et pour la même vue fashion, j'ai la même vue, la même policier, donc je suis stap. Nous a égal B 3 et et exemple donc on va s'arrêter à V 2 en pratique on va dire Sélestat. Bref. Que TAC TAC TAC. Voilà donc là, vous avez un premier votre premier algorithme de de de R félicitations, vous pouvez implémenter le policier utilisation. Voilà, voilà. Tac TAC TAC. Je. Achille, tu m'as pas aussi, je suis pas sûr. Universel, 18 pas. Euh. Ça c'est un peu ce qu'on me dérangerait. Peut-être après la pause, mais du coup ? Ouais. Ouais, ouais, très bien, on va faire une pause maintenant. Et ouais, on se retrouve. Dans 15 Min. Du coup ceux qui sont en ligne ça va ? Vous êtes pas trop loin normalement et si vous avez des questions attendant bah hésitez pas. Excusez-moi Monsieur, Ouais, j'ai oublié de signer s'il vous plaît. Y a possibilité de me mettre réseau s'il vous plaît. Ce que je. Je vais pas envoyer un code normalement si mais en fait au bout de 15 min on peut plus mettre le code. En pratique ça c'est un peu iznaga parce que si ton paramètre est trop petit, trop grand, et cetera, t'auras le même problème. Du coup on le fait pas juste on le fait pas et on passe à quelque chose de plus efficace qui est de directement. Directement ? Faire de la value visage. Je suppose en à tout moment que tu utilises la politique optimale dans la politique. Pardon l'automne pour la value function que tu as. Là si vous remarquez tu vois on update bien à chaque chaque nouvelle itération on update bien notre value function. Avec. Le maximum que je puisse avoir. 2. Le maximum sur les actions, donc en supposant. Tu vois, tu te retrouves dans time value fonction OK ? Je regarde tes États autour étudier. Ok, si jamais je fais cette action là, ma value Fund ce sera ça. Si je regarde à côté. Ahh le maximum ce sera ça. Ok pas ouf, je regarde à côté. Ah là et là une bonne fonction donc je prends celle-là. Donc implicitement ce que t'as dit c'est. Je suppose que ma policie en fait, c'était d'aller dans la direction où j'ai la meilleure valeur aussi. Donc c'est un peu comme si implicitement ta policier était toujours la policier optimale associée à la fonction que t'as ensemble. Et donc j'ai plus besoin de passer par PI. Là ici, on a plus de notion de PI. En fait, je veux directement, je converge vers la fonction tout de suite. Dire qu'on va redescendre à 012 guitare exactement, là on va converger directement vers 0-1-2. 3. Je veux avoir ma connexion. On continue peut-être, ça donnera plus d'éléments, mais. Mais c'est c'est très proche de l'idée qu'on avait là tu vois ici OK on est avec une policy, une value function qui est nulle au début. Sauf que du coup la première itération c'est je remarque que. En fait, la case d'à côté. J'ai une récompense de un parce que j'ai terminé la viande OK et donc ma valeur était richon ça prend le maximum. Ça fait si ma police était bien d'aller sur cette pièce là. Eh Ben off. Je peux pas après hop maintenant ma policier. Il m'a valu fonction, je prends le maximum possible donc ça va être 9. La chaleur parce que là faut que je j'étais là, je vais pas juste avant, je me dis c'est intelligence ça ? Mais mais en gros, et on parle de la fin pour rejoindre exactement et tu vas de la même manière qu'on de ce dont on parlait hier avec Belman tu vas tu vas propager tes informations. Donc là petit détail je suis désolé aussi, c'est pas les mêmes les mêmes graphiques, mais ici on est dans un labyrinthe qui est un peu différent de celui d'hier où on a juste une récompense de un quand on est arrivé et 0 sinon. Ok. Et donc c'est pour ça que là on a on a 0 9 du coup et pas et pas 0-1-2-3 donc là je vais me servir même. Ouais en gros moi là je vais partir hein. De toute façon quand je suis au-dessus je vois descendre un c'est ça ? Je découvre en œuvre la seule chose qui exactement. Et bah du coup j'ai une nouvelle valeur de belle fonction. OK donc c'est c'est comme si à chaque fois je supposais que ma policier était optimal pour la la validation que. Ça c'est le meilleur exactement. Et donc je suppose que je vais finir par faire ce coup là. On peut le faire là ? Carmouze, donc ça soit le programme et attends donc enfin on pourrait aussi le faire pour les programmes parce qu'en soit oui en soit la probabilité elle est là hein. Oui mais si on se retrouve avec bah du coup c'est moins certain avec des programmes. Ouais ouais non ça ça va, mais beaucoup plus de temps combien là là on ça a beaucoup plus de voilà et du coup ça donne ça vraiment je propage petit à petit jusqu'à arriver à une nouvelle fonction, qui va qui, qui va nous donner. Et en pratique on verra que c'est la la value itérations et qu'on verre directement vers Vista. Là c'est là, c'est des stars, ouais je crois. Ouais. Donc là on a, on a rien représenté techniquement, on a juste modélisé V. On a trouvé la vallée sanction optimale et donc bien entendu, à partir de ça, je peux en extraire la policier optimale. Hop. Alors ça ? On revient sur la police, direction. Donc là on est on est sur une value function qui est pas encore complète. Et du coup ? Là, on va changer. Là on est sur la improvement. Ok, là on vient, on vient de d'évaluer la policy. Ok, on a la policier, c'est les flèches qu'on a, OK ? Et là je viens d'évaluer vos policiers, OK ? Donc j'ai trouvé les bonnes valeurs jusqu'ici parce que j'avais la bonne police juste ici. Par contre ici. On retrouve avec 0 parce qu'en fait notre notre agent il arrive jamais il fait que des il fait que des tours sur lui-même il est perdu. Donc là j'ai pris ma balle, son chat et donc je vais faire un Trojan sur ma sur ma policier. Donc là typiquement on a on on va arrêter de tourner à droite à ce moment-là mais on va aller vers une direction. Mais en pratique de en Python, ce sera dominatrice. Des valeurs ? Bah non, en pratique on fait ça pour trouver les valeurs. Là moi je te donne un je te donne un labyrinthe avec des points noirs et enfin directement à circuler et moi je te la. La question que je pose c'est si je mets un agent dans mon labyrinthe, comment est-ce qu'il en soit ? On les donne avec un seul et c'est avec la récompense que tu après la récompense elle est un peu dure, il y a juste un tout à la fin. Donc là c'est on continue notre notre itération, on vient de changer notre policier, hop la nouvelle value fonction associée à notre policier, elle est là. Donc là on regarde, on regarde, y a encore un nouveau progrès à faire à cet endroit là. Hop, on change les 2 cas, on calcule une nouvelle nouvelle nouvelle fonction. Là pareil on a encore un. Un update à faire sur notre policier of et donc là on se rend compte que là on a un chemin qui. On a un chemin qui va dans la bonne direction, enfin qui qui termine, sauf que il est pas optimal parce que. Parce qu'on pourrait aller plus vite. Donc là on remarque bien que cette polyéthylène permet bien de changer de décision à cet endroit là pour aller dans la bonne direction ici. Et in fine pareil pour la dernière pour arriver à la fameuse value function terminal associée à la police. Ce que c'est ? La Chine demande. La fonction c'est quoi cette définition ? Comment ? En espérance, car il a pas bien appris entendre dire toi et ce qui peut venir, les questions de la jeune. C'est la seule des rewards. Référence de la. Espérance. Non ? Ta chambre. Pas parce que c'est quand même pas compliqué. C'est sur votre fiche que vous avez que je vous ai mis sur moi. Et donc ça en pratique c'est y a quoi ? Si on développe, c'est l'espérance de la somme des tout. Les gamins. Et. De 0 à T. Sachant si. Égal à 6. Il manque quelque chose sachant ? Bah avec les actions non non alors attendez. Un gamin exposant. OK. C'est l'espérance de mon gars, c'est l'espérance de la somme qui ont géré donc de récompense. Donc en pratique oui. Ça donne un très bon indice sur la récompense qu'on va avoir, mais c'est pas exactement ça, pardon ? C'est l'espérance de la somme de. C'est quand même. En paramètres, c'est une récompense et. La révolution, elle prend en paramètre l'État, c'est tout, elle te donne à tel État. En supposant que tu as une politique PI définie, c'est vrai que j'aurais dû, j'aurais dû. C'est quoi l'espérance de mon gars ? Ok Voilà ok, est-ce que pour les autres c'est clair du coup ce qu'on fait là les 2° qu'on viendra ? La politisation, ou. Je calcule 20 un challenge, je modifie ma policy, j'y terre et cetera versus la value itération directement. Hein je supp ? C'est toujours une politique qui est optimale. Et je la modifie en temps réel pour pour propager mes valeurs. C'est que la, pour tout le monde ou pas. Oui, c'est clair. Oui. Top. Eh Ben on va pouvoir. Euh, on va pouvoir passer le TP alors ? Euh là, je vous ai mis des petits. Des petits tableaux récapitulatifs. Pour voilà, pour vous donner une idée de OK, ce qui se passe, c'est quoi notre, notre objectif, ce qu'on fait donc là dans la policier ? Évaluation. On calcule la value flashing associée à Py. De manière itérative. La police évaluation, ça nous permet d'implémenter l'algorithme de Polytech Sion, où cette fois on évalue la policy, puis. On en déduit une nouvelle policière un peu meilleure, et cetera. Et enfin la vallée, Itération où on s'extrait de la nécessité de modéliser pied. On suppose toujours que puis il est-il est crédit et dans on trouve des stars directement. Et. Pas de la validation t'as pas besoin de ski avant la value itération c'est directement. Donc c'est un. Par contre, dans le cas de la politique itération, la besoin de la politique, la police, évaluation pour, pour, pour. Ok. Euh, TAC, TAC, TAC. Ouais donc ça pareil, c'est plutôt orienté contrôle pour ceux qui pour ceux qui ont. Ouais, je sais pas, je sais pas trop quoi dire dessus, c'est la manière, la manière de voir les choses quoi. Voilà. Tac TAC TAC ouais, rapide rap de de ce qu'on vient de voir. Euh, l'approbation dynamique, là ça va. C'est hyper efficace pour ce qu'on fait, donc c'est à dire pour des labyrinthes pas trop gros. N'empêche que. On va être limité quand on va avoir des grippes trop grosses, mais donc en pratique ce que je vais vous reteniez si un jour vous avez à choisir entre ce genre de modèle ? Si si vous avez une grille, vous savez que vous avez capacité de modéliser tout en même temps. Vous aurez pas mieux que le Danemark proving pour pour résoudre le problème. Et vous savez, vous trouverez exactement la meilleure solution en attendant en un temps fini. OK, ça sert à rien de faire des modèles de type des réseaux de neurones et cetera sur un mails. Vous pouvez vous amuser à faire, mais vous allez dépenser beaucoup d'énergie pour rien alors que alors que value itération en. Attend ridicule, il va voir le problème de. Ça se passe-passe. Ouais. Ouais, ouais, c'est ça, on a besoin pour ça de connaître le environnement en entier. Et et et la prochaine étape du cours ça va être de voir bah comment est-ce qu'on fait pour quand on a pas le monde quoi, quand on n'a pas accès à tout ce qui se passe dans l'environnement et quand notre notre agent il a uniquement une observation, donc une observation, vous vous rappelez ce que c'est hein ? J'en ai parlé dans hier simplement pour l'instant on n'a pas eu d'observation parce qu'en fait on. Explore uniquement l'État ? Quoi, on n'a pas besoin. C'est ça, oui. C'est ça ? Alors qu'un robot en soi ou même nous. Quand on fait enfin ce qu'on observe du monde réel, c'est vraiment une infime partie de sa complexité. Et et en temps réel. Nous, on est toujours en train de faire des prédictions sur qu'est ce qui va se passer ? Je vois quelqu'un marcher de la porte d'un probablement qui va se mettre à ouvrir la porte, en faire du bruit donc. Donc. Ok et bien c'est c'est le moment, on va pouvoir s'amuser un petit peu. YATP, qui est sur le moodle. Euh. Je vous laisse le télécharger, l'ouvrir. Si vous voulez des plages sympas, il va falloir que vous l'ouvriez en local. Parce que sur collab, les plots ne marchent pas. Donc il y a pas besoin de GPU ou quoi ? Donc non, vous pouvez tous faire tourner ça en local. Faut juste. Il faut juste un environnement qu'on a avec Jupiter. Je sais pas si ça parle à tout le monde. Si vous avez des soucis bah vous me vous me demandez. Je peux essayer de vous aider un peu mais donc dans l'idée vous avez juste. Enfin, vous avez tous déjà fait des Jupiter dans le bocal local ou il y en a qui ont jamais fait ça ? Bon très bien et n'hésitez pas à vous aider entre vous, c'est des soucis. Là je pense que je vous laisse, je vous laisse 10 Min pour mettre en place votre environnement, et cetera. Et ensuite on lance le TP pour ceux qui ont besoin de des questions à poser. Voilà, essayez de le lire normalement il est très détaillé déjà, donc vous avez même pas besoin de moi pour pour le pour le compléter, mais simplement je reste. Je reste là pour répondre à vos questions si jamais j'ai besoin. Monsieur, oui le TP moi sur moodle je le vois en format seul. Non, normalement. Comme ça en fait le les INB c'est essaie d'utilisation donc c'est normal mais normalement tu peux l'ouvrir quand même avec Jupiter. Je vais vous mettre sur sur teams aussi au cas où, mais normalement normalement c'est censé marcher. Oui, c'est bon autant pour moi. Ok. Bon, je vais me chercher un café, j'ai des problèmes pour demander à l'heure.